Estadística

Ángel A. Juan
Blanca de la Fuente
Alicia Vila

PID_00159944

Material docente de la UOC

Universitat Oberta de Catalunya
www.uoc.edu
FONDO SOCIAL EUROPEO
MINISTERIO DE INDUSTRIA, TURISMO Y COMERCIO
plan avanza
alquimia



Ángel A. Juan

Licenciado en Matemáticas por la Universidad de Valencia, Máster en Tecnologías de la Información por la UOC y Doctor en Matemática Computacional Aplicada por la UNED. En la actualidad es profesor agregado de Estadística y Simulación en los Estudios de Informática, Multimedia y Telecomunicación de la UOC. Asimismo, es profesor asociado de Estadística Aplicada en la Universidad Politécnica de Cataluña. Sus líneas de investigación se centran en los ámbitos de la simulación por computador, el análisis de datos y el aprendizaje de las matemáticas en entornos en línea, ámbitos en los que ha publicado numerosos artículos en revistas y libros internacionales. Para más información, podéis consultar http://ajuanp.wordpress.com



Blanca de la Fuente

Doctora en Ciencias Biológicas (1988) por la Universidad Complutense de Madrid desde 1988. Profesora del Departamento de Estadística e Investigación Operativa II (Métodos de Decisión) de la Facultad de Ciencias Económicas y Empresariales de la Universidad Complutense de Madrid y Consultora de la Universitat Oberta de Catalunya. Ha sido docente desde 1992 de asignaturas del área de la Estadística en diversas titulaciones de universidades públicas y privadas. Sus áreas de investigación son el análisis multivariante y aplicaciones de nuevas metodologías docentes en la enseñanza universitaria.



Alicia Vila

Licenciada en Matemáticas por la Universidad de Valencia. Profesora de ciclos formativos en el ámbito de la informática, en particular en los campos de programación y bases de datos. Ha impartido docencia en el área de Probabilidad y Estadística en diferentes titulaciones de la Universitat Oberta de Catalunya.

El encargo y la creación de este material docente han sido coordinados por el profesor: Víctor Cavaller (2011)

El proyecto E-ALQUIMIA ha sido apoyado por el Ministerio de Industria, Turismo y Comercio en el marco de las ayudas para la realización de actuaciones sobre contenidos digitales en el marco del Plan Avanza, y por la Unión Europea a través de los Fondos Comunitarios. Referencia: PAV-10000-2007-275

Primera edición: febrero 2011

Av. Tibidabo, 39-43, 08035 Barcelona
Realización editorial: Eureca Media, SL
Diseño: Manel Andreu
Depósito legal: B-1.339-2010
ISBN: 978-84-693-9717-6

Introducción

La asignatura de *Estadística* está dirigida a los estudiantes del grado de Información y Documentación.

Los estudios de Información y Documentación ofrecen múltiples salidas profesionales desde el trabajo en centros de información (bibliotecas, mediatecas, centros de documentación, archivos), hasta la gestión de información en organizaciones del sector privado o público (análisis de la información, gestión documental, gestión de contenidos, arquitectura de la información, webmaster) y la gestión de sistemas de información.

En general, la estadística se ha convertido en una herramienta imprescindible en el campo de las ciencias sociales, en los trabajos de investigación y a la hora de desarrollar profesionalmente tareas relacionadas con la gestión, la interpretación de datos y la toma de decisiones.

En el marco concreto de las competencias que tiene que desarrollar un gestor de la información y de la documentación, la estadística es un instrumento muy útil, sea cual sea el campo profesional que se quiere desarrollar.

Estos materiales introducen los conceptos estadísticos más necesarios para su formación, utilizando un enfoque práctico y aplicado. En este sentido, se da prioridad a la adquisición de conceptos y métodos aplicados, evitando el uso de un excesivo formalismo matemático. A priori, no se necesitan conocimientos previos de estadística, ya que esta asignatura se tratará desde cero y suponiendo que el estudiante no ha trabajado nunca en este campo.

El material didáctico está constituido por cinco módulos:

1. Estadística descriptiva, que incluye una introducción a la estadística y a la descripción de datos mediante tablas, gráficos y estadísticos, así como al concepto de probabilidad y de distribución de probabilidad.

2. Inferencia de información para una población, que incluye distribuciones, intervalos y contrastes.

3. Inferencia de información para dos poblaciones, sobre los contrastes de hipótesis para dos poblaciones.

4. Relación entre variables: causalidad, correlación y regresión, que incluye modelos de regresión simple (lineales, cuadráticos y cúbicos).

5. Introducción al diseño y análisis de encuestas, sobre las aplicaciones estadísticas a la selección de muestras y al análisis de cuestionarios.

Objetivos

El objetivo fundamental es introducir al estudiante en el uso de la metodología estadística para describir y compilar datos, construir muestras aleatorias válidas, comprobar hipótesis y elaborar modelos estadísticos.

A grandes rasgos, las competencias que se pretenden alcanzar son:

1. Entender la importancia de la estadística en la sociedad moderna.

2. Aprender a organizar y resumir de forma descriptiva un conjunto de datos de una muestra mediante gráficos, tablas de frecuencias y estadísticos.

3. Comprender el concepto de probabilidad de un acontecimiento y descubrir sus principales propiedades y aplicaciones.

4. Conocer las principales distribuciones estadísticas que se usan para modelar el comportamiento de variables discretas y continuas, y utilizarlas en pruebas de hipótesis.

5. Aplicar e interpretar la inferencia estadística en poblaciones.

6. Entender la importancia de las encuestas y los cuestionarios en la sociedad de la información y conocer su elaboración y aplicación.

7. Aprender a usar software estadístico y de análisis de datos como instrumento básico en la aplicación práctica de los conceptos y las técnicas estadísticas.

Contenidos

Bibliografía

Anderson, D.; Sweeney, D.; Williams, T. (2008). *Statistics for Business and Economics*. South-Western College Pub. ISBN: 0324658370.

Berk, K.; Carey, P. (2003). *Data Analysis with Microsoft Excel*. Duxbury Press. ISBN: 0534407145.

Bowermann, B. L.; O'Connell, R. T. (1997). *Applied Statistics: Improving Business Processes*. Irwin. ISBN: 025819386X.

Draper, N. R.; Smith, H. (1998). *Applied Regression Analysis. Wiley*. ISBN: 0471170828.

Fowler, F. (2008). *Survey Research Methods*. Sage Publications, Inc. ISBN: 1412958415.

Johnson, R.; Kuby, P. (2006). *Elementary Statistics*. Duxbury Press. ISBN: 0495017639.

Lohr, S. (1999). *Sampling: Design and Analysis*. Duxbury Press. ISBN: 0534353614.

Moore, D. (2006). *The Basic Practice of Statistics*. W. H. Freeman. ISBN: 071677478X.

Moore, D.; McCabe, G. (2005). *Introduction to the Practice of Statistics*. W. H. Freeman. ISBN: 0716764008.

Myer, R. H. (1990). *Classical and Modern Regression with Applications*. PWS. ISBN: 0534921787.

Rea, L.; Parker, R. (2005). *Designing and Conducting Survey Research: A Comprehensive Guide*. Jossey Bass. ISBN: 078797546X.

Ryan, B.; Joiner, B.; Cryer, J. (2005). *MINITAB Handbook. Brooks/Cole - Thomson Learning Inc*. ISBN: 0534496008.

Settle, R.; Alreck, P. (2003). *Survey Research Handbook*. McGraw-Hill/Irwin. ISBN: 0072945486.

Thompson, S. (2002). *Sampling*. Wiley-Interscience. ISBN: 0471291161.

Estadística descriptiva univariante

Modelos estadísticos para la descripción de datos univariantes

Alicia Vila y Ángel A. Juan

PID_00161058

UOC
Universitat Oberta de Catalunya
www.uoc.edu

Índice

Introducción

Las sociedades modernas son ricas en datos: la prensa escrita, la televisión y la radio, Internet y las intranets de las organizaciones ofrecen cantidades inmensas de datos que pueden ser procesados y analizados. Esto convierte a la estadística en una ciencia interesante y útil puesto que proporciona estrategias y herramientas que permiten obtener información a partir de dichos datos. Además, gracias a la evolución de la tecnología (ordenadores y software estadístico) hoy en día es posible automatizar gran parte de los cálculos matemáticos asociados al uso de técnicas estadísticas, lo que permite extender su uso a un gran rango de profesionales en ámbitos tan diversos como la biología, las ciencias empresariales, la sociología o las ciencias de la información.

La práctica de la estadística requiere aprender a obtener y explorar los datos –tanto numéricamente como mediante gráficos–, a pensar sobre el contexto de los datos y el diseño del estudio que los ha generado, a considerar la posible influencia de observaciones anómalas en los resultados obtenidos, a discutir la legitimidad de los supuestos requeridos por cada técnica y, finalmente, a validar la fiabilidad de las conclusiones derivadas del análisis. La estadística requiere tanto de conocimientos sobre los conceptos y técnicas empleados como de la suficiente capacidad crítica que permita evaluar la conveniencia de usar unas u otras técnicas según el tipo de datos disponible y el tipo de información que se desea obtener.

En este módulo inicial de la asignatura, se examinan los datos procedentes de una única variable: en primer lugar se explica cómo organizar y resumir dichos datos, tanto numérica como gráficamente (estadística descriptiva); en segundo lugar, se introducen los conceptos básicos asociados con la idea de probabilidad; finalmente, se presentan algunos modelos matemáticos que permiten analizar el comportamiento de algunas variables.

Objetivos

Los objetivos académicos que se plantean en este módulo son los siguientes:

1. Entender la importancia de la estadística en la sociedad moderna.

2. Aprender a organizar y resumir un conjunto de datos procedentes de una variable mediante gráficos, tablas de frecuencias y estadísticos descriptivos.

3. Comprender el concepto de probabilidad de un suceso y descubrir sus principales propiedades y aplicaciones.

4. Conocer las principales distribuciones estadísticas que se usan para modelar el comportamiento de variables discretas y continuas.

5. Saber calcular probabilidades asociadas a cada una de las distribuciones introducidas.

6. Aprender a usar software estadístico o de análisis de datos como instrumento básico en la aplicación práctica de los conceptos y técnicas estadísticas.

1. Introducción a la Estadística

La Estadística es la ciencia que se ocupa de obtener datos y procesarlos para transformarlos en información. Es, por tanto, un lenguaje universal ampliamente utilizado en las ciencias sociales, en las ciencias experimentales, en las ciencias de la salud y en las ingenierías. Las Tecnologías de la Información y la Comunicación (TIC) han incrementado notablemente la producción, diseminación y tratamiento de la información estadística. En particular, Internet es una fuente inagotable de datos que pueden ofrecer información y, a partir de ella, conocimiento. Por otra parte, la constante evolución de los ordenadores personales y de los **programas informáticos de estadística** y análisis de datos posibilita y facilita el análisis de grandes cantidades de datos mediante el uso de técnicas estadísticas y de minería de datos. En la Sociedad de la Información se hace pues imprescindible disponer de un cierto conocimiento estadístico incluso para poder comprender e interpretar correctamente los indicadores económicos (IPC, inflación, tasa de desempleo, Euribor, etc.), los indicadores bibliométricos (factor de impacto de una revista, cuartil en el que se sitúa, vida media de las citas recibidas, etc.) o los indicadores sociales (esperanza de vida, índice de alfabetización, índice de pobreza, indicador social de desarrollo sostenible, etc.) a los que frecuentemente se hace referencia en los medios de comunicación.

Nota

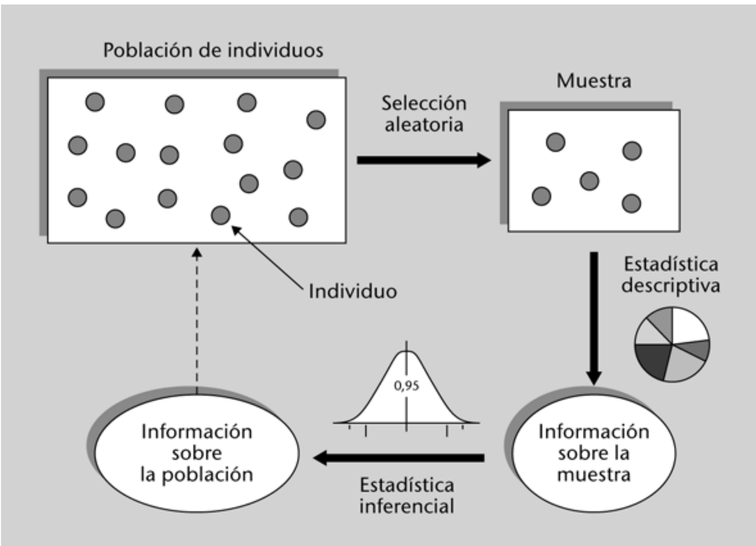
Las agencias gubernamentales, como el Instituto Nacional de Estadística (INE) o el Eurostat proporcionan datos sobre casi cualquier ámbito socioeconómico.

Software estadístico

En la actualidad existen excelentes **programas informáticos** para el análisis estadístico de datos. Algunos ejemplos son: MINITAB, SPSS, MS Excel, SAS, R, S-Plus, Statgraphics o Statistica.

El campo de la Estadística se puede dividir en dos grandes áreas: la estadística descriptiva y la estadística inferencial (figura 1).

Figura 1. Estadística descriptiva y estadística inferencial



La estadística descriptiva se ocupa de la obtención, presentación y descripción de datos procedentes de una muestra o subconjunto de una población de individuos. Por su parte, la estadística inferencial usa los resultados obtenidos

mediante la aplicación de las técnicas descriptivas a una muestra para inferir información sobre el total de la población a la que pertenece dicha muestra.

Algunos términos básicos

A lo largo de este material se usarán abundantes términos estadísticos, muchos de ellos bastante conocidos. A continuación se presentan y revisan algunos de estos términos básicos que conviene entender bien:

- **Población:** colección o conjunto de elementos (individuos, objetos o sucesos) cuyas propiedades se desean analizar. Ejemplos: (a) los estudiantes universitarios de un país; (b) el conjunto de periódicos en Internet; (c) el conjunto de revistas indexadas en el Science Citation Index (SCI), etc.

- **Muestra:** cualquier subconjunto de elementos de la población. Ejemplos: (a) los estudiantes de una determinada universidad; (b) los periódicos en línea centrados en aspectos económicos; (c) las revistas indexadas en el SCI de una determinada editorial, etc.

- **Muestra aleatoria:** muestra cuyos elementos han sido escogidos de forma aleatoria. Ejemplos: (a) un subconjunto de doscientos estudiantes escogidos al azar (mediante el uso de números aleatorios) de entre todos los matriculados en universidades de un país; (b) un subconjunto de cincuenta periódicos en línea escogidos al azar; (c) un subconjunto de quince revistas indexadas en el SCI escogidas al azar, etc.

- **Marco del muestreo:** lista que contiene aquellos elementos de la población candidatos a ser seleccionados en la fase de muestreo. No necesariamente coincidirá con toda la población de interés, ya que en ocasiones no será posible identificar a todos los elementos de la población. Ejemplos: (a) lista de todos los estudiantes matriculados en universidades de un país en un semestre concreto; (b) relación de periódicos en línea disponibles en un momento dado; (c) lista de todas las revistas indexadas en el SCI en un año específico, etc.

- **Variable aleatoria:** característica de interés asociada a cada uno de los elementos de la población o muestra considerada. Ejemplos: (a) la edad de cada estudiante; (b) el número de visitas diarias que recibe cada periódico en línea; (c) el factor de impacto de cada revista, etc.

- **Datos u observaciones:** conjunto de valores obtenidos para la variable de interés en cada uno de los elementos de la muestra. Ejemplos: (a) las edades registradas son {25, 23, 19, 28…}; (b) las visitas diarias registradas son {1326, 1792, 578, 982…}; (c) los factores de impacto registrados son {2,3; 1,7; 8,2…}.

- **Experimento:** estudio en la que el investigador controla o modifica expresamente las condiciones del mismo con la finalidad de analizar los distin-

tos patrones de respuesta en las observaciones. Ejemplos: (a) estudiar cómo varían las calificaciones de un grupo de estudiantes según dispongan o no de ordenadores con acceso a Internet en las aulas; (b) estudiar cómo varía el número de visitas a un periódico en línea según se opte o no por incluir noticias sensacionalistas en su portada; (c) estudiar cómo varía el factor de impacto de un grupo de revistas según éstas se incluyan o no en una base de datos de reconocido prestigio, etc.

- **Inspección o encuesta:** estudio en el que el investigador no pretende modificar las condiciones de la muestra con respecto a la variable de interés sino simplemente obtener los datos correspondientes a unas condiciones estándar. Ejemplos: (a) registrar las calificaciones de los estudiantes de un máster determinado; (b) realizar una encuesta a los lectores de un periódico en línea; (c) obtener el factor de impacto asociado a cada una de las revistas de una muestra, etc.

- **Parámetro:** valor numérico que sintetiza alguna propiedad determinada de la población. Los parámetros se asocian a toda la población y suelen representarse con letras del alfabeto griego como μ (mu), σ (sigma), etc. Ejemplos: (a) la edad media de todos los estudiantes universitarios de un país; (b) el número máximo de visitas diarias recibido por algún periódico en línea; (c) el rango o diferencia entre el mayor y el menor factor de impacto del conjunto de revistas indexadas en el SCI, etc.

- **Estadístico:** valor numérico que sintetiza alguna propiedad determinada de una muestra. Los estadísticos se asocian a una muestra y se suelen representar por letras del alfabeto latino como $\overline{x}$, s, etc. Ejemplos: (a) la edad media de los estudiantes de una muestra aleatoria; (b) el número máximo de visitas diarias recibidas por algún periódico deportivo en línea; (c) el rango o diferencia entre el mayor y el menor factor de impacto de las revistas de una editorial, etc.

- **Variable cualitativa o categórica:** variable que categoriza o describe cualitativamente un elemento de la población. Suele ser de tipo alfanumérico, pero incluso en el caso en que sea numérica no tiene sentido usarla en operaciones aritméticas. Ejemplos: (a) el teléfono o el correo electrónico de un estudiante; (b) la dirección IP de un periódico en línea; (c) el ISSN de una revista, etc.

- **Variable cuantitativa o numérica:** variable que cuantifica alguna propiedad de un elemento de la población. Es posible realizar operaciones aritméticas con ella. Ejemplos: (a) el importe de la beca que recibe un estudiante; (b) los ingresos que genera un periódico en línea; (c) el número de revistas publicadas por una editorial, etc.

- **Variable cuantitativa discreta:** variable cuantitativa que puede tomar un número finito o contable de valores distintos. Ejemplos: (a) edad de un es-

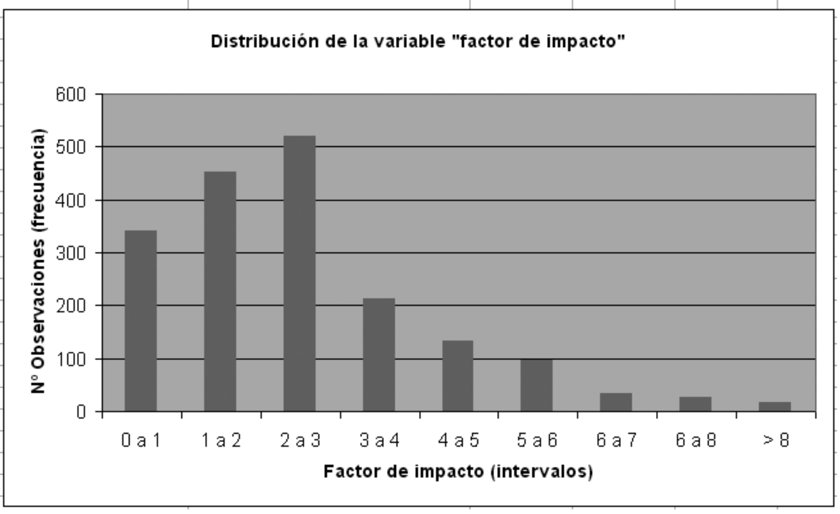
tudiante; (b) número de enlaces a otras fuentes de información que ofrece un periódico en línea; (c) calificación que obtiene una revista en una escala entera de 1 a 5, etc.

- **Variable cuantitativa continua:** variable cuantitativa que puede tomar un número infinito (no contable) de valores distintos. Ejemplos: (a) altura o peso de un estudiante; (b) tiempo que transcurre entre la publicación de una encuesta en línea y el instante en que ya la han completado un centenar de internautas; (c) factor de impacto (sin redondear) de una revista, etc.

- **Distribución de una variable:** en sentido amplio, una distribución es una tabla, gráfico o función matemática que explica cómo se comportan o distribuyen los valores de una variable, es decir, qué valores toma la variable así como la frecuencia de aparición de cada uno de ellos. Ejemplo: dada una muestra aleatoria de revistas, la distribución de la variable "factor de impacto de una revista" puede representarse mediante una tabla de frecuencias o mediante una gráfica como se aprecia en la figura 2. Se observa que trescientas cuarenta y dos de las revistas consideradas tienen un factor de impacto entre 0 y 1, cuatrocientas cincuenta y dos de las revistas tienen un factor de impacto entre 1 y 2, etc.

Figura 2. Distribución de una variable aleatoria

Factor de impacto	N.º de observaciones (frecuencia)
0 a 1	342
1 a 2	452
2 a 3	521
3 a 4	212
4 a 5	132
5 a 6	97
6 a 7	35
6 a 8	27
> 8	16



2. Descripción de datos mediante tablas y gráficos

Cuando se dispone de un conjunto de observaciones procedentes de una muestra conviene hacer un primer análisis exploratorio de éstas mediante gráficos y tablas que ayuden a interpretar los datos y a extraer información de los mismos. Existen diferentes tipos de gráficos que pueden usarse en esta fase exploratoria y el uso de unos u otros dependerá en gran medida del tipo de datos de los que se disponga (cualitativos o cuantitativos), así como de la información que se desee visualizar. En este apartado se presentaran algunos de los gráficos y tablas más habituales para la descripción de **datos univariantes**.

Datos univariantes

Los datos univariantes son los que provienen de una única variable. En algunos casos, los datos pueden proceder de dos o más variables y, entonces, se usa la expresión bivariante (si se trata de dos variables) o multivariante (si se consideran más de dos).

Gráficos y tablas para datos cualitativos o categóricos

Si se dispone de datos cualitativos o categóricos, pueden sintetizarse mediante una tabla que recoja, para cada categoría: el número de veces que aparece (frecuencia absoluta), el porcentaje de apariciones sobre el total de observaciones (frecuencia relativa), así como los acumulados de ambos valores. La tabla 1 muestra esta información para la variable "número de *hotspots* (conexiones *wi-fi*) identificados en cada comunidad autónoma".

Tabla 1. Ejemplo de tabla de frecuencias para una variable categórica

Comunidad autónoma	***Hotspots* por comunidad autónoma**			
	Frecuencia	**Frecuencia acumulada**	**Frecuencia relativa**	**Frec. rel. acumulada**
Andalucía	885	885	11,9%	11,9%
Aragón	177	1.062	2,4%	14,2%
Asturias	148	1.210	2,0%	16,2%
Cantabria	164	1.374	2,2%	18,4%
Castilla-La Mancha	144	1.518	1,9%	20,3%
Castilla y León	302	1.820	4,0%	24,4%
Cataluña	1.391	3.211	18,6%	43,0%
C. Valenciana	622	3.833	8,3%	51,3%
Extremadura	137	3.970	1,8%	53,2%
Galicia	516	4.486	6,9%	60,1%
I. Baleares	183	4.669	2,5%	62,5%
I. Canarias	151	4.820	2,0%	64,6%
La Rioja	126	4.946	1,7%	66,3%
Madrid	1.776	6.722	23,8%	90,0%
Murcia	160	6.882	2,1%	92,2%
Navarra	153	7.035	2,0%	94,2%
País Vasco	430	7.465	5,8%	100,0%
Totales	**7.465**		**100,0%**	

Nota

Observad que la **frecuencia acumulada** se obtiene sólo con ir acumulando frecuencias anteriores.

Además de mediante una tabla de frecuencias, suele ser habitual representar datos categóricos mediante el uso de gráficos circulares (figura 3) o bien mediante diagramas de barras (figura 4).

Figura 3. Ejemplo de gráfico circular para una variable categórica

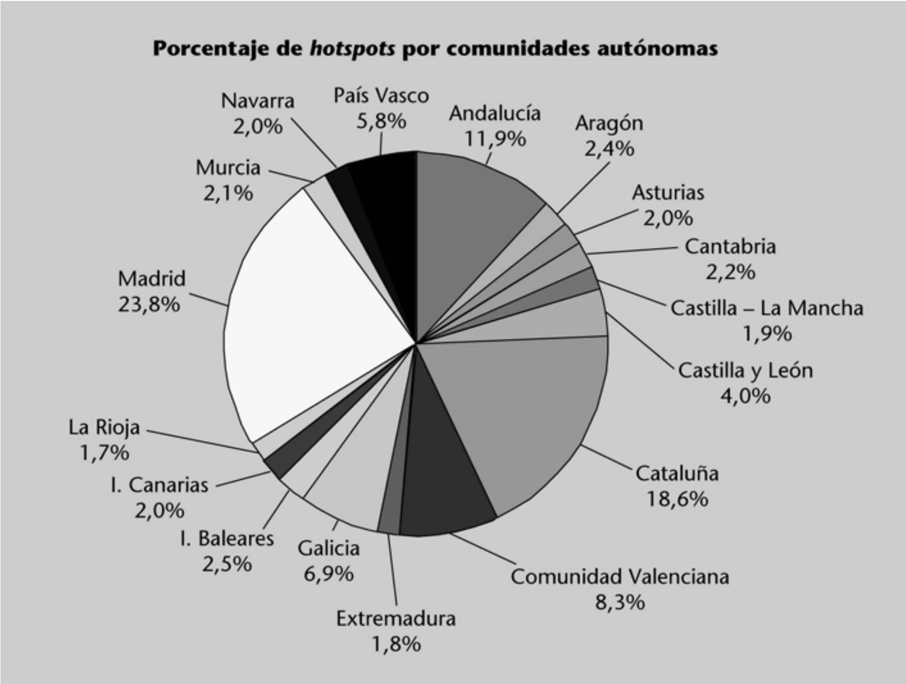
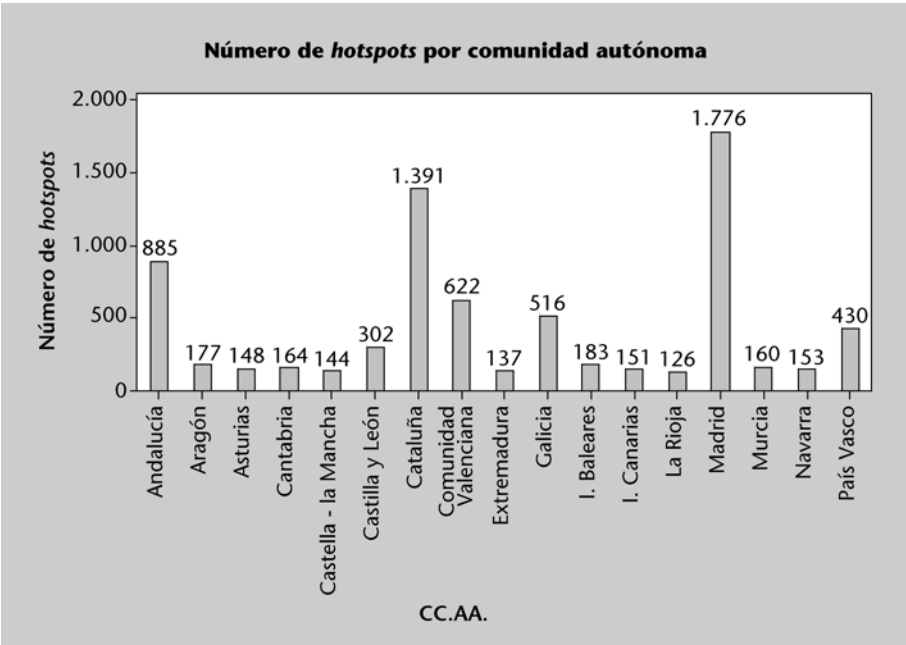
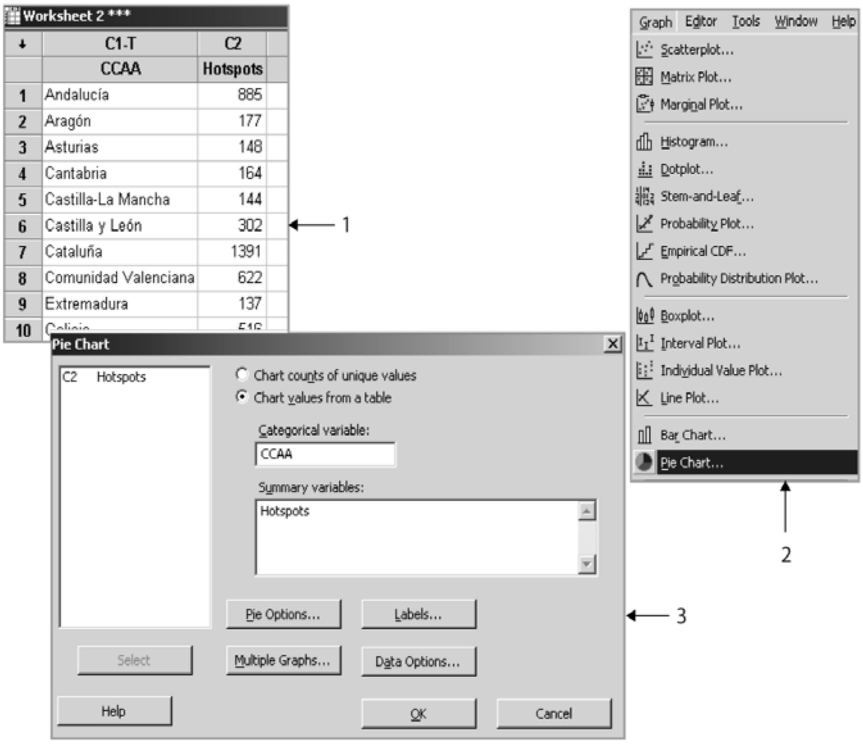


Figura 4. Ejemplo de diagrama de barras para una variable categórica



Este tipo de gráficos pueden crearse fácilmente con cualquier programa estadístico o de análisis de datos (p. ej.: Minitab, MS Excel, SPSS, etc.). La figura 5 muestra los pasos básicos para generar un gráfico circular (*pie chart*) con Minitab. La generación de un diagrama de barras (*bar chart*) se consigue de forma similar, al igual que ocurre con la mayoría de los gráficos que se presentan en este apartado.

Figura 5. Pasos a seguir para la generación de un gráfico circular con Minitab



Un gráfico que también suele usarse bastante para describir datos cualitativos es el llamado diagrama de Pareto. Este gráfico está compuesto por: (a) un diagrama de barras en el que las categorías están ordenadas de mayor a menor frecuencia y (b) una línea que representa la frecuencia relativa acumulada (figura 6).

Pasos a seguir

Una vez introducidos los datos en el programa *(1)*, se sigue la ruta ***Graph*** > ***Pie Chart (2)*** y se seleccionan las variables en la ventana correspondiente *(3)*.

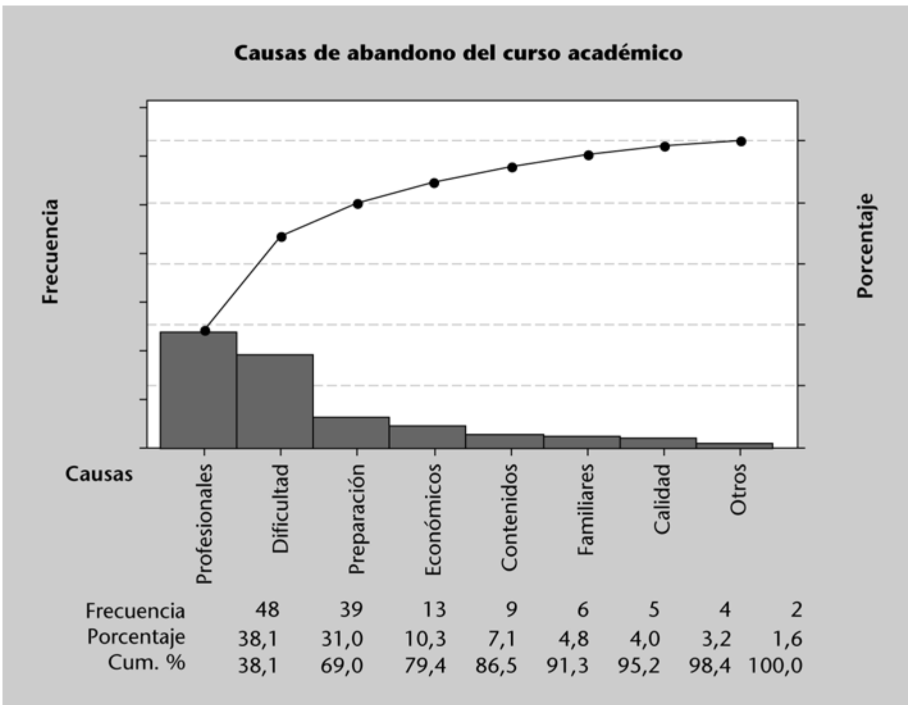
Nota

Las capturas de pantalla de Minitab corresponden a la **versión 15** de este programa. Es posible que otras versiones ofrezcan ligeras diferencias en los menús y ventanas, aunque básicamente el proceso será el mismo. Para obtener más detalles sobre las opciones disponibles, siempre es posible consultar la ayuda en línea del programa o bien alguno de los numerosos manuales de uso que se pueden encontrar en Internet.

Diagrama de Pareto

Para generar un diagrama de Pareto en Minitab hay que usar la ruta *Stat* > *Quality Tools*.

Figura 6. Diagrama de Pareto sobre las causas de abandono de un curso



Los diagramas de Pareto son muy útiles para detectar cuándo un porcentaje reducido de categorías (p. ej.: un 20% de las categorías) “acapara” o representa un porcentaje alto de observaciones (p. ej.: un 80% de los datos). Estos fenómenos de excesiva representatividad por parte de unas pocas categorías suelen darse con frecuencia en contextos socioeconómicos (p. ej.: un porcentaje reducido de los ciudadanos de un país acapara un alto porcentaje de la renta), educativos (p. ej.: un porcentaje reducido de causas generan la mayor parte de los abandonos del curso) o de ingeniería de la calidad (p. ej.: un alto porcentaje de fallos son debidos a un número muy reducido de causas). Identificar aquellas pocas categorías que representan una gran parte del porcentaje total puede servir para corroborar ciertos desequilibrios distributivos –como una distribución poco equilibrada de las rentas en un país o de los sueldos en una empresa–, o para proporcionar pistas sobre los principales factores de causa de un problema –como el alto nivel de abandono de un curso o un elevado nivel de fallos en un servicio o producto–.

Gráficos y tablas para datos cuantitativos

En el caso de datos cuantitativos, su representación gráfica o mediante tablas permite apreciar la forma de su distribución estadística, es decir, la forma en que se comporta la variable de interés (cuáles son los valores medios o centrales, cuáles son los valores más habituales, cómo varía, cómo de dispersos son los valores, si muestra algún patrón de comportamiento especial, etc.).

Uno de los gráficos más sencillos de elaborar es el llamado gráfico de puntos (*dotplot*). Se trata de un gráfico en el que cada punto representa una o más observaciones. Los puntos se apilan uno sobre otro cuando se repiten los valores observados (figura 7).

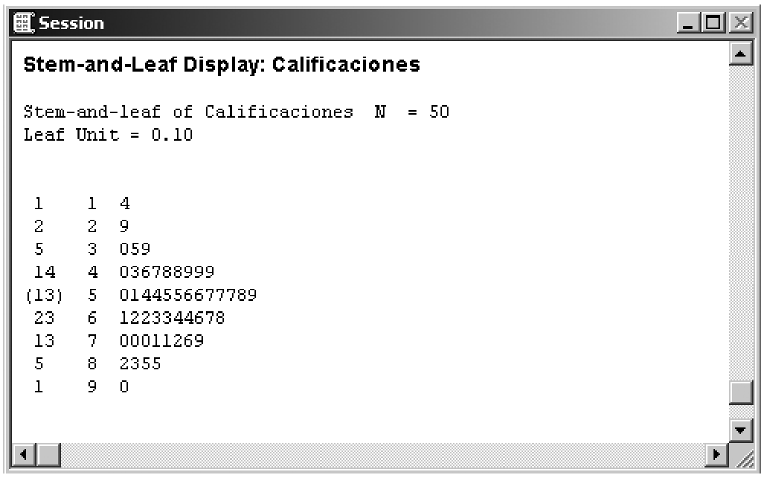
Figura 7. Gráfico de puntos para las calificaciones de un curso



Un gráfico similar, aunque algo más elaborado y con una orientación transpuesta de los ejes, es el llamado diagrama de tallos y hojas (*stem-and-leaf*). En él también se representan los valores observados pero usando los propios valores numéricos en lugar de puntos, lo que proporciona un mayor nivel de detalle. La figura 8 muestra un ejemplo de gráfico de tallos y hojas para los mismos datos empleados en la figura 7. Se observa que el gráfico se ha construido a partir de una muestra de cincuenta calificaciones y que

se ha usado una unidad de hoja (*leaf*) de 0,1. Esto significa que la segunda columna del gráfico representa la parte entera de la calificación, mientras que cada uno de los números situados a su derecha representa la parte decimal de una observación con dicha parte entera. Así, se pueden leer las siguientes calificaciones por orden de menor a mayor: 1,4, 2,9, 3,0, 3,5, 3,9, 4,0, 4,3, etc.

Figura 8. Gráfico de hojas y tallos para las calificaciones de un curso



> **Atención**
>
> Cabe destacar que en un gráfico de tallos y hojas los datos se apilan de izquierda a derecha en lugar de arriba abajo como ocurre con el gráfico de puntos.

Cuando las observaciones generan un número elevado de valores distintos, resulta recomendable agruparlos en clases o intervalos disjuntos de igual tamaño. De ese modo, cada observación se clasifica en una clase o intervalo según su valor. La tabla 2 muestra un ejemplo de tabla de frecuencias en el que se han agrupado los datos en intervalos. La frecuencia de cada intervalo viene determinada por el número de observaciones cuyos valores están en dicho intervalo. La marca de clase representa el valor medio del intervalo.

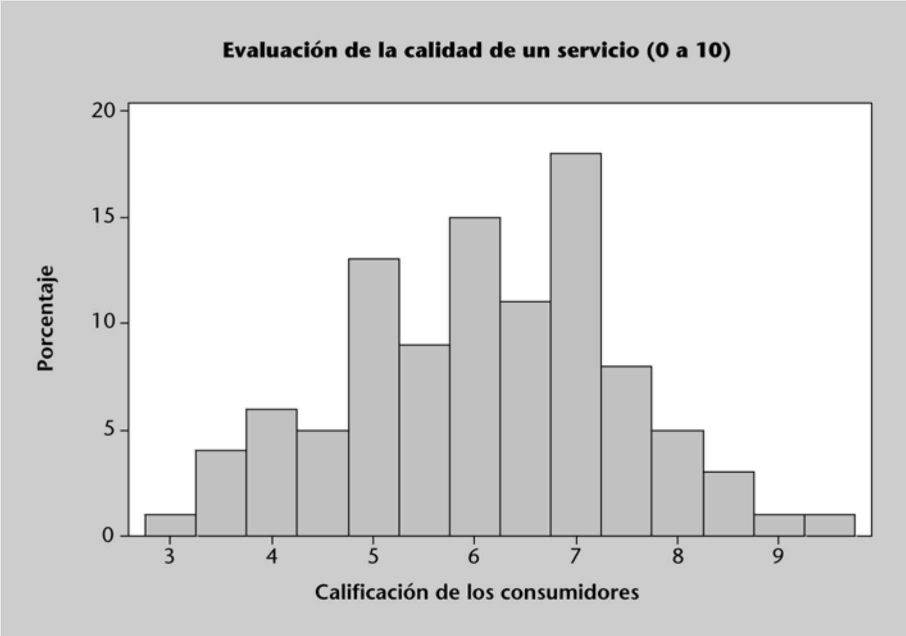
Tabla 2. Ejemplo de tabla de frecuencias agrupadas usando intervalos

Intervalo	Marca de clase	Frecuencia	Frecuencia relativa
[0, 2)	1	12	8,1%
[2, 4)	3	23	15,5%
[4, 6)	5	67	45,3%
[6, 8)	7	31	20,9%
[8, 10)	9	15	10,1%
Totales		**148**	**100,0%**

Un gráfico que utiliza también intervalos para agrupar los datos a representar es el histograma. El histograma muestra la frecuencia (absoluta o relativa) de cada clase, lo que permite visualizar de forma aproximada la distribución de los datos (figura 9). Sin embargo, hay que tener presente que la forma final del histograma puede variar bastante según el número de intervalos que se definan para agrupar los datos, lo que a veces no permite apreciar correctamente la forma exacta de la distribución estadística que siguen las observaciones.

> **Nota**
>
> Una regla habitual es definir $\sqrt{n}$ clases o intervalos, siendo *n* el número de observaciones disponibles.

Figura 9. Histograma de una distribución aproximadamente normal



La figura 9 muestra un histograma con forma de campana: es una forma bastante simétrica, que presenta una mayor altura en la parte central y disminuye paulatinamente en las "colas" o extremos. Esta forma es bastante habitual y suele caracterizar el comportamiento de muchas variables (p. ej.: notas numéricas en un examen, peso o altura de individuos, temperaturas diarias, etc.). Sin embargo, también es habitual encontrarse con variables que muestran patrones de comportamientos completamente distintos. Por ejemplo, la figura 10 muestra un histograma en el que se aprecia una distribución más "uniforme" u homogénea de los datos, mientras que la figura 11 muestra un histograma en el que se aprecia una distribución asimétrica o "sesgada" de los mismos.

Figura 10. Histograma de una distribución aproximadamente uniforme

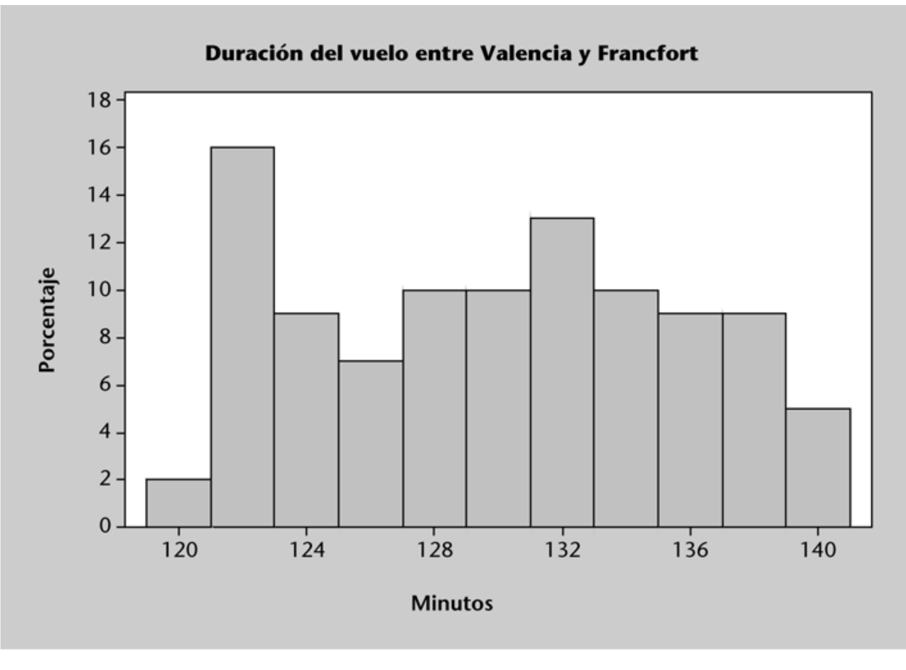
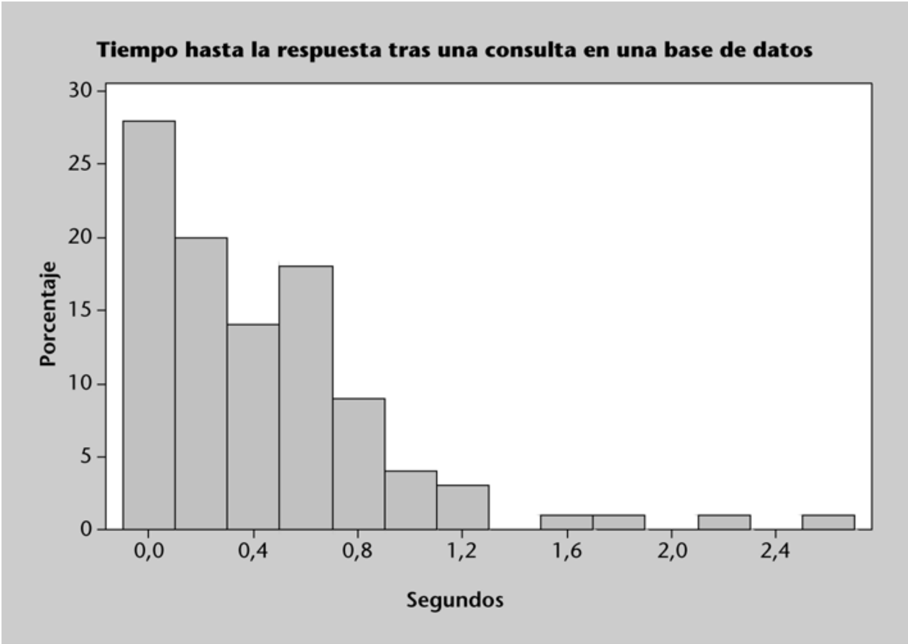


Figura 11. Histograma de una distribución sesgada a la derecha



3. Descripción de datos mediante estadísticos

Dado un conjunto de n datos u observaciones, x_1, x_2, ..., x_n, asociadas a una variable de interés X, suele ser útil sintetizar algunas de sus principales propiedades en unos pocos valores numéricos. Los estadísticos descriptivos son, precisamente, estos valores numéricos capaces de proporcionar información a partir del conjunto de las observaciones. Estos estadísticos resultan muy útiles a la hora de entender el comportamiento de los datos, ya que un simple valor numérico es capaz de describir propiedades tan relevantes como, por ejemplo, el valor promedio del conjunto de datos, el valor máximo, el valor mínimo, el valor que se repite con más frecuencia, un índice de dispersión o variabilidad, etc.

Web

Recordar que la World Wide Web (p. ej., Wikipedia, etc.) es una excelente fuente de consulta para ampliar los conceptos y definiciones estadísticas que se proporcionan en este y otros módulos. Un recurso especialmente interesante, por cuanto ofrece una visión muy completa de conceptos y técnicas estadísticas, es el libro en línea de StatSoft http://www.statsoft.com/textbook/.

Como ya se comentó anteriormente, estos estadísticos hacen referencia a una muestra de observaciones y suelen representarse mediante letras del alfabeto latino ($\overline{x}$, s, etc.), lo que permite distinguirlos claramente de sus parámetros asociados que sintetizan propiedades de toda la población y se representan mediante letras griegas (μ, σ, etc.). Básicamente pueden distinguirse dos grupos de estadísticos descriptivos: (a) los de centralización, que proporcionan información sobre cuáles son los valores "centrales" del conjunto de datos (p. ej.: el valor promedio de los datos) y (b) los de dispersión, que explican cómo se sitúan y varían los datos con respecto a los valores "centrales" (p. ej.: el rango o diferencia entre el valor máximo y el valor mínimo de los datos).

Nota

Recordar que los símbolos μ y σ se pronuncian como "mu" y "sigma", respectivamente. La pronunciación de otros símbolos del alfabeto griego se puede consultar, p. ej., en Wikepedia.

Estadísticos de centralización

A continuación se presentan los estadísticos de centralización más usados habitualmente:

- **Media (*mean*)**: la media (también conocida por valor promedio o valor esperado) de un conjunto de observaciones muestrales se representa con el símbolo $\overline{x}$. Intuitivamente, la media simboliza el "centro de masas" o "punto de equilibrio central" del conjunto de datos considerado. El parámetro asociado, la media poblacional, se representa por μ. Para calcular la media de un conjunto de datos se usa la siguiente expresión:

$$\overline{x} = \frac{x_1 + x_2 + ... + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

  Ejemplo: la media de los cinco datos siguientes {6, 3, 8, 6, 4} es

$$\overline{x} = \frac{6+3+8+6+4}{5} = \frac{27}{5} = 5{,}4$$

Media muestral

Recordar que la media muestral es un ***estadístico*** que hace referencia al "centro de masas" de los datos de una muestra (subconjunto de la población), mientras que la media poblacional es un **parámetro** que representa el "centro de masas" de toda la población.

- **Mediana (*median*):** la mediana de un conjunto de observaciones muestrales suele representarse con el símbolo $\tilde{x}$. En el caso de una población, el

parámetro mediana se denota con M. Una vez se ordenan todos los datos de menor a mayor, la mediana es aquel valor que deja a su izquierda la mitad de las observaciones (es decir, es aquel valor tal que el número de observaciones más pequeñas que él coincide con el número de observaciones mayores que él). Los pasos para calcular la mediana son: (1) ordenar los datos de menor a mayor, (2) calcular la posición i que ocupa la mediana en el conjunto ordenado de datos, $i = \frac{n+1}{2}$ y (3) seleccionar la observación x_i (la que ocupa la posición determinada en el paso anterior). Cabe observar que si el número de datos n es impar (p. ej.: $n = 5$), la posición i será un valor entero (p. ej.: $i = 3$) que corresponderá con un valor concreto, x_i, del conjunto de datos. Sin embargo, si n es par (p. ej.: $n = 6$), la posición i será un número no entero (p. ej.: $i = 3{,}5$), en cuyo caso la mediana vendrá dada por el promedio de los dos valores que ocupan las posiciones enteras más cercanas a i (en este caso por el promedio de los valores que ocupan las posiciones 3 y 4).

Ejemplo: dado el conjunto de ocho datos {5, 11, 7, 8, 10, 9, 6, 9}, lo primero es ordenarlos de menor a mayor, con lo que se obtiene la serie {5, 6, 7, 8, 9, 9, 10, 11}; ahora, la posición de la mediana vendrá dada por $i = \frac{8+1}{2} = 4{,}5$, es decir, la mediana estará entre los valores que ocupan las posiciones 4 y 5, por lo que se calcula el promedio de ambos para dar el valor de la mediana, es decir: $\tilde{x} = \frac{8+9}{2} = 8{,}5$.

Es importante destacar que la media es muy sensible a la existencia de valores extremos (*outliers*), es decir, la inclusión o no de un valor que esté muy alejado del resto de los datos puede cambiar considerablemente el valor resultante de la media. Por el contrario, la mediana se ve mucho menos afectada por la presencia de dichos valores, lo que significa que la mediana es un "centro" más estable que la media en el sentido de que se ve menos afectado por la presencia de valores extremos en los datos.

- **Moda (*mode*):** la moda de un conjunto de datos es el valor que más veces se repite (el de mayor frecuencia).

  Ejemplo: la moda de la serie de datos {6, 3, 4, 8, 9, 6, 6, 3, 4} es 6, puesto que es el valor que más veces aparece en la serie.

Estadísticos de dispersión

Se presentan ahora los principales estadísticos de dispersión que, como se ha comentado anteriormente, proporcionan información sobre la variabilidad del conjunto de datos:

- **Rango (*range*):** el rango de un conjunto de datos es la diferencia entre el valor máximo y el mínimo de los mismos.

  Ejemplo: dado el conjunto de datos {2, 3, 8, 3, 5, 1, –8}, su rango es 8 – (–8) = 16

- **Varianza muestral (*sample variance*):** la varianza de una muestra se representa por el símbolo s^2. En el caso de una población, el parámetro varianza se representa con el símbolo σ^2. La varianza muestral será mayor cuanto mayor sean las diferencias entre cada una de las observaciones x_i y la media de los datos $\overline{x}$, en concreto:

$$s^2 = \frac{(x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + \ldots + (x_n - \overline{x})^2}{n-1} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

  Esto significa que la varianza es una medida de la dispersión de los datos con respecto a su media, es decir, cuando menor sea la varianza, tanto más agrupados estarán los datos alrededor de su valor promedio. Por el contrario, cuanto mayor sea la varianza, tanto más dispersos estarán los datos.

  Ejemplo: la varianza muestral de la serie de 5 datos {6, 3, 8, 5, 3} es:

$$s^2 = \frac{(6-5)^2 + (3-5)^2 + (8-5)^2 + (5-5)^2 + (3-5)^2}{5-1} = 4,5$$

- **Desviación estándar (*standard deviation*):** la desviación estándar (o típica) de una muestra se representa con el símbolo *s*, mientras que la desviación estándar de una población se representa con σ. La desviación estándar es la raíz cuadrada positiva de la varianza, esto es: $s = \sqrt{s^2}$ (o, dicho de otro modo, la varianza es el cuadrado de la desviación estándar).

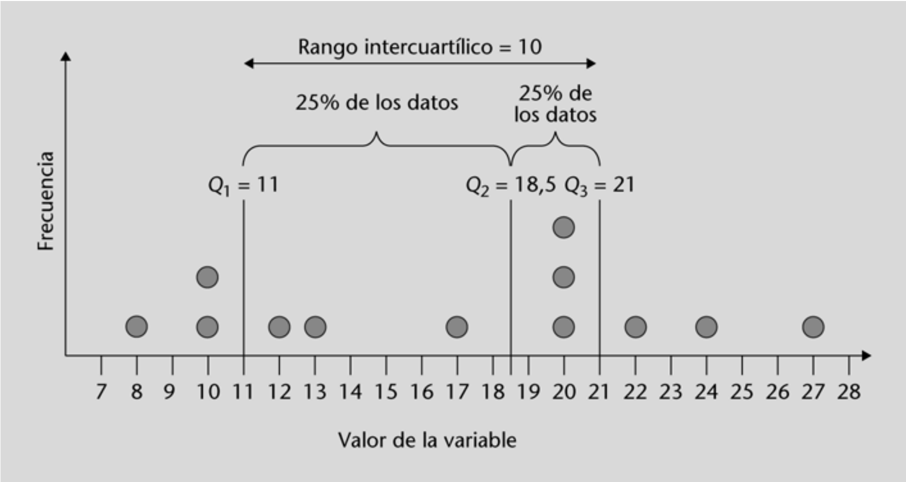
  Ejemplo: para los datos del ejemplo anterior, $s = \sqrt{4,5} = 2,1$

  Al igual que ocurría con la varianza, a mayor desviación estándar más dispersión en los datos y viceversa.

- **Cuartiles (*quartiles*):** en un conjunto de *n* observaciones ordenadas de menor a mayor valor, se pueden considerar tres valores numéricos concretos llamados cuartiles que dividen el conjunto en cuatro partes, cada una de ellas conteniendo una cuarta parte de las observaciones (figura 12). El primer cuartil, Q_1, es el valor que deja la cuarta parte de los datos ordenados a su izquierda (es decir, un 25% de los datos muestran valores inferiores a él y un 75% de los datos muestran valores superiores a él). Por su parte, el segundo cuartil, Q_2, es aquel valor que deja la mitad de los datos ordenados a su izquierda (es decir, un 50% de los datos muestran valores inferiores a él y un 50% de los datos muestran valores superiores a él). Finalmente, el tercer cuartil, Q_3, es aquel va-

lor que deja tres cuartas partes de los datos ordenados a su izquierda (es decir, un 75% de los datos muestran valores inferiores a él y un 25% de los datos muestran valores superiores a él).

Figura 12. Cuartiles de un conjunto ordenado de datos



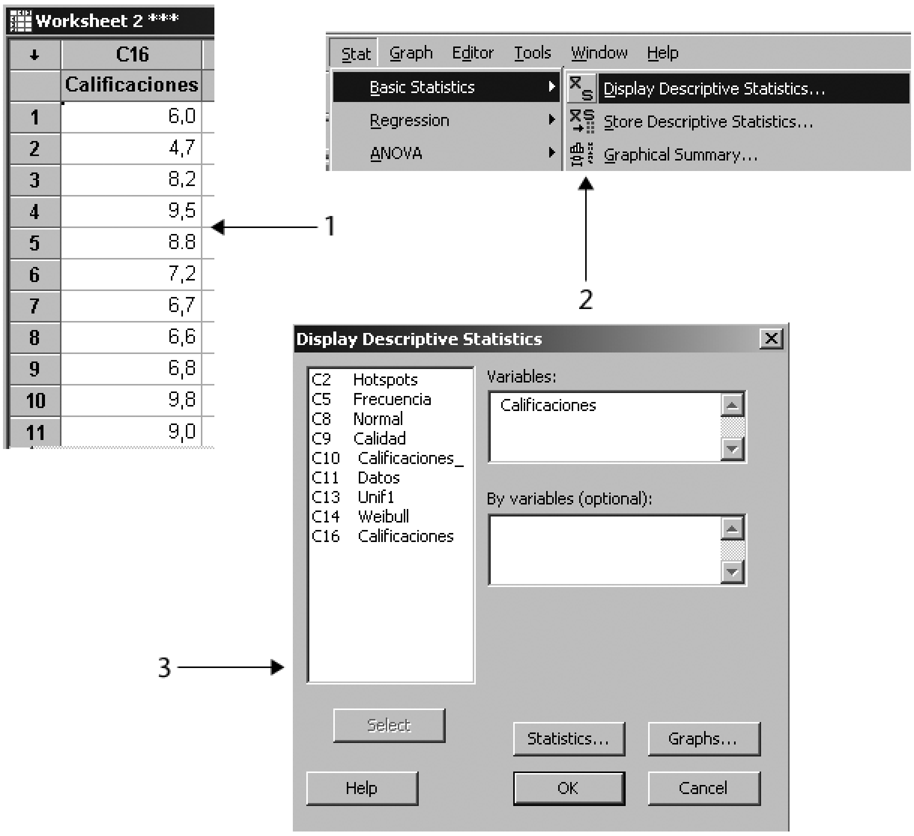
Obsérvese que, en realidad, el cuartil segundo o Q_2 coincide con el concepto de mediana presentado anteriormente. Los cuartiles son muy útiles a la hora de clasificar una observación en una determinada franja del conjunto de datos, por ejemplo, si la observación es inferior a Q_1 significa que ésta se encuentra situada entre el 25% de valores más bajos; si la observación es superior a Q_3 significa que está situada entre el 25% de valores más altos, etc.

- **Rango intercuartílico (*inter-quartilic range*):** este rango suele representarse como *IQR* y es simplemente la diferencia entre el tercer cuartil y el primer cuartil, es decir: $IQR = Q_3 - Q_1$. El rango intercuartílico indica el espacio que ocupan el 50% de las observaciones "centrales" (figura 12), por lo que, de forma similar a lo que ocurría con la varianza, da una medida de la dispersión de los datos (a mayor *IQR* mayor dispersión y viceversa).

Obtención de estadísticos descriptivos mediante programas informáticos

En la práctica, es habitual utilizar algún programa estadístico o de análisis de datos para calcular los estadísticos anteriores e incluso algunos estadísticos adicionales que proporcionen información sobre el conjunto de datos. En la figura 13 se muestran los pasos básicos necesarios para obtener los principales estadísticos descriptivos con Minitab. El *output* del programa, para un ejemplo con cincuenta observaciones, se muestra en la figura 14. Por su parte, la figura 15 muestra una serie de estadísticos descriptivos generados con MS Excel para el mismo conjunto de datos (en este caso los cuartiles se han obtenido usando las fórmulas integradas de Excel).

Figura 13. Pasos para calcular estadísticos descriptivos con Minitab



Pasos a seguir

Una vez introducidos los datos en el programa *(1)*, se sigue la ruta ***Stat*** > ***Basic Statistics*** > ***Display Descriptive Statistics…*** ***(2)*** y se seleccionan las variables en la ventana correspondiente *(3)*.

Figura 14. Estadísticos descriptivos obtenidos con Minitab

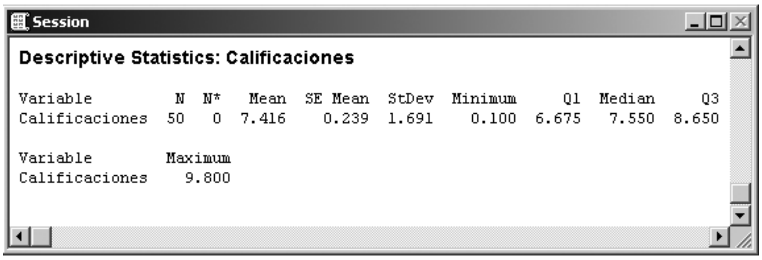


Figura 15. Estadísticos descriptivos calculados con Excel

	A	B	C	D
1	Calificaciones			
2	6,0		*Calificaciones*	
3	4,7			
4	8,2		Media	7,416
5	9,5		Error típico	0,239205488
6	8,8		Mediana	7,55
7	7,2		Moda	7,2
8	6,7		Desviación estándar	1,691438224
9	6,6		Varianza de la muestra	2,860963265
10	6,8		Curtosis	5,922785035
11	9,8		Coeficiente de asimetría	-1,720324942
12	9,0		Rango	9,7
13	7,7		Mínimo	0,1
14	8,6		Máximo	9,8
15	5,8		Suma	370,8
16	6,4		Cuenta	50
17	9,5			
18	7,4		Cuartil primero	6,725
19	7,2		Cuartil segundo	7,55
20	8,8		Cuartil tercero	8,6
21	7,4			

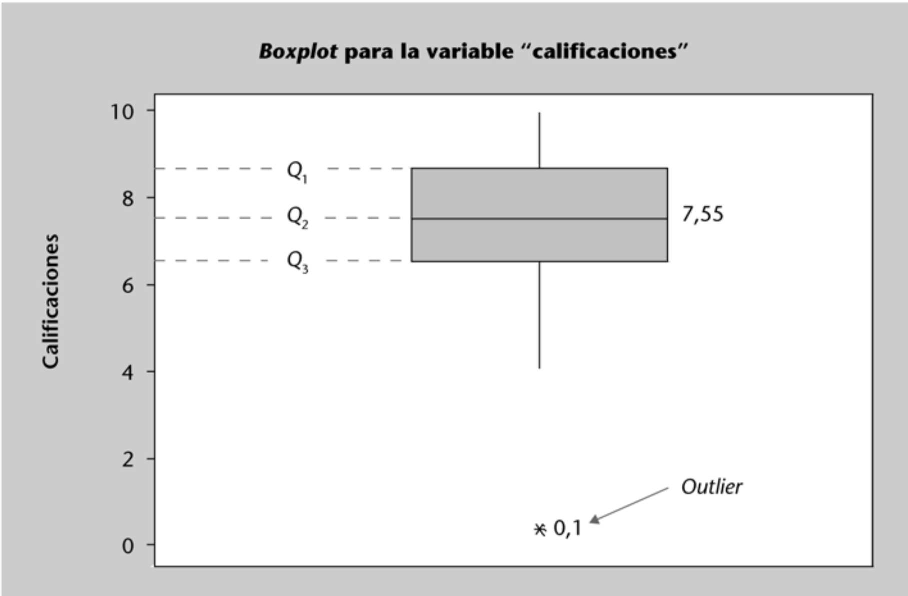
Diferencias en los métodos de cálculos

Cabe destacar que hay ligeras diferencias entre los valores de los cuartiles calculados por Minitab y los correspondientes valores de Excel. Ello se debe a que usan métodos de cálculo distintos. Una discusión interesante sobre los diferentes métodos existentes para calcular los cuartiles se puede encontrar en: http://mathforum.org/library/drmath/view/60969.html.

Diagrama de cajas y bigotes (*boxplot*)

Usando los cuartiles es posible construir un tipo de gráfico, el diagrama de cajas y bigotes (*boxplot*), que resulta muy útil para visualizar la distribución de los datos. Este diagrama está compuesto por una caja central, definida por los cuartiles primero y tercero, que contiene el 50% "central" de las observaciones, y dos segmentos situados en los respectivos extremos de la caja, representando cada uno de ellos el 25% de las observaciones extremas (figura 16).

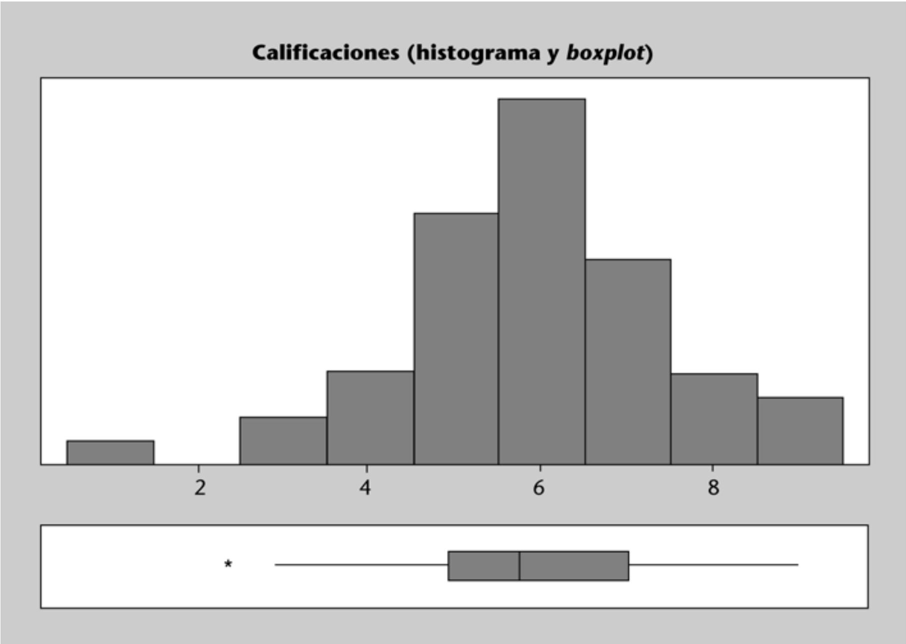
Figura 16. Diagrama de cajas y bigotes (*boxplot*) y valores extremos (*outliers*)



El diagrama de cajas y bigotes sirve también para identificar posibles valores anómalos (*outliers*), que se encuentran excesivamente alejados del resto de los datos, es decir: o bien son extremadamente grandes o bien extremadamente pequeños en comparación con el resto de observaciones. Estos valores anómalos se suelen representar mediante un asterisco, y pueden ser debidos a un error en el registro de los datos o bien a valores que, en realidad, se encuentran extremadamente alejados del resto de observaciones (p. ej.: el precio de un Ferrari cuando se compara con precios de turismos de gama media). Identificar valores anómalos en un conjunto de observaciones es importante, puesto que el análisis de los datos puede dar resultados muy distintos en función de que se consideren o no dichos valores en el estudio (por ejemplo, la media y la varianza de un conjunto de datos pueden cambiar de forma notable según se incluya o no uno de estos valores extremos).

La estrecha relación existente entre el histograma y el *boxplot* se puede observar en la figura 17. En cierto sentido, el *boxplot* se puede interpretar como un histograma visto desde arriba. En este caso, la zona del *boxplot* situada entre los cuartiles primero y tercero correspondería a la zona central del histograma. Además, en ambos casos queda identificado el valor anómalo (*outlier*) así como la forma aproximadamente simétrica del resto de la distribución.

Figura 17. Relación entre histograma y *boxplot*



4. El concepto de probabilidad

Un **experimento aleatorio** es aquel en el que no es posible conocer a priori el suceso resultante que acontecerá pero, sin embargo, sí es posible observar un cierto patrón regular en los resultados que van sucediendo cuando el experimento se repite muchas veces. Por ejemplo, cuando se considera el experimento aleatorio consistente en lanzar una moneda (o un dado) al aire, no es posible predecir cuál será el **suceso resultante** del experimento, es decir, si saldrá cara o cruz (o qué número saldrá en el caso del dado); sin embargo, sí se puede afirmar que tras muchos lanzamientos el porcentaje o proporción de sucesos "cara" obtenidos será muy próximo al 50% o 1/2 (en el caso del dado, el porcentaje o proporción de sucesos "3" obtenidos será muy próximo a 0,1667 o 1/6). Este porcentaje o proporción de aparición de un suceso tras muchas repeticiones del experimento es lo que da lugar a la idea de probabilidad:

Ejemplo

La **probabilidad** de un suceso es siempre un número entre 0 y 1. Así, por ejemplo, una probabilidad de 0,25 representa un porcentaje de aparición del 25% o, equivalentemente, una proporción de 1/4.

Se define la **probabilidad de un suceso** A, $P(A)$, como el porcentaje o proporción de aparición de dicho suceso en una serie extraordinariamente larga de repeticiones del experimento, todas ellas independientes entre sí.

El requisito de independencia entre las distintas repeticiones del experimento aleatorio significa que el resultado de cada repetición del experimento no está condicionado por los resultados obtenidos en repeticiones anteriores (p. ej.: cuando se lanza varias veces una moneda al aire, el suceso resultante de cada nuevo lanzamiento es independiente de los resultados obtenidos en lanzamientos previos).

Ejemplo 1 de probabilidades

En el experimento "lanzamiento de una moneda al aire", es posible considerar los siguientes sucesos o potenciales resultados: C = {cara}, X = {cruz}, Ω = {cara o cruz} y $\varnothing$ = {ni cara ni cruz}. Los dos últimos sucesos se conocen, respectivamente, como suceso seguro Ω (que incluye todos los resultados posibles) y suceso imposible o conjunto vacío $\varnothing$ (que no incluye ningún resultado derivado de la ejecución del experimento). En este caso, parece claro que $P(C) = 0,5$ (es decir, si se repitiera el experimento muchas veces, aproximadamente el 50% de las mismas serían caras), $P(X) = 0,5$, $P(\Omega) = 1$ (es decir, en el 100% de los lanzamientos saldrá o bien cara o bien cruz) y $P(\varnothing) = 0$ (es decir, en el 0% de los lanzamientos no se obtendrá resultado alguno).

Ejemplo 2 de probabilidades

En el experimento aleatorio "lanzamiento de un dado", es posible considerar sucesos o potenciales resultados como los siguientes: {1}, {2}, {3}, {4}, {5}, {6},

Ω = {un número entre 1 y 6}, $\varnothing$ = {ningún número entre 1 y 6}. En este caso, $P(\{1\}) = 1/6$ (tras muchas repeticiones, uno de cada seis lanzamientos acabará siendo un 1), $P(\{2\}) = 1/6$, $P(\{3\}) = 1/6$, $P(\{4\}) = 1/6$, $P(\{5\}) = 1/6$, $P(\{6\}) = 1/6$, $P(\Omega) = 1$ y $P(\varnothing) = 0$.

Observar, además, que también es posible considerar sucesos compuestos como, por ejemplo, par = {2, 4, 6}, impar = {1, 3, 5}, mayor2 = {3, 4, 5, 6}, menor3 = {1, 2}, etc. En este caso, *P*(par) = 3/6 = 1/2, *P*(impar) = 1/2, *P*(mayor2) = 4/6 = 2/3, *P*(menor3) = 2/6 = 1/3.

Propiedades básicas de las probabilidades

Hay una serie de propiedades básicas que debe satisfacer cualquier probabilidad. Estas propiedades son muy útiles a la hora de calcular probabilidades de sucesos complejos a partir de probabilidades ya conocidas o fáciles de obtener:

1) La probabilidad de cualquier suceso A siempre es un número situado entre 0 y 1 (ambos inclusive), es decir $0 \leq P(A) \leq 1$.

Ejemplo: en los ejemplos anteriores, todas las probabilidades halladas eran valores entre 0 y 1.

2) La probabilidad del suceso imposible o conjunto vacío $\varnothing$ es siempre 0, es decir, $P(\varnothing) = 0$. En otras palabras, cuando se hace un experimento aleatorio siempre se obtiene algún resultado y, por tanto, la proporción de "no-resultados" es 0.

Ejemplo: en los ejemplos anteriores, $P(\varnothing) = 0$.

3) La suma de las probabilidades de todos los posibles resultados del experimento aleatorio siempre vale 1. En otras palabras, la probabilidad del suceso seguro es siempre 1.

Ejemplo: En el ejemplo de la moneda, $P(\Omega) = 1 = P(C) + P(X)$; en el ejemplo del dado, $P(\Omega) = 1 = P(\{1\}) + P(\{2\}) + P(\{3\}) + P(\{4\}) + P(\{5\}) + P(\{6\})$.

4) La probabilidad de que un suceso no ocurra es 1 menos la probabilidad de que sí ocurra, es decir: $P(no\ A) = 1 - P(A)$.

Ejemplo: en el ejemplo de la moneda, $P(C) = 0{,}5 = 1 - P(no\ C) = 1 - P(X)$; en el ejemplo del dado, *P*(par) = 0,5 = 1 – *P*(no par) = 1 – *P*(impar); $P(\varnothing) = 1 - P(\Omega)$.

5) Si dos sucesos *A* y *B* no tienen resultados comunes (son disjuntos), la probabilidad de que ocurra $A \cup B$ es la suma de las probabilidades, es decir, si *A* y *B* son disjuntos, $P(A \cup B) = P(A) + P(B)$.

Ejemplo: en el ejemplo de la moneda, $P(C \cup X) = P(C) + P(X) = 1$; en el ejemplo del dado, $P(\{1, 2\}) = P(\{1\}) + P(\{2\}) = 2/6 = 1/3$; $P(\Omega \cup \emptyset) = P(\Omega) + P(\emptyset) = 1 + 0 = 1$.

6) En general, para cualesquiera dos sucesos A y B se cumplirá que $P(A \cup B) = P(A) + P(B) - P(A \cap B)$, donde "$A \cap B$" es el conjunto de posibles resultados que satisfacen los sucesos A y B a la vez. Hay que tener en cuenta que cuando A y B son disjuntos (no tienen resultados en común), "$A \cap B$" = $\emptyset$ y, por tanto, $P(A \cup B) = P(A) + P(B) - P(\emptyset) = P(A) + P(B) - 0 = P(A) + P(B)$, que es la expresión vista en la propiedad anterior.

Ejemplo: en el ejemplo del dado, $P(\text{par} \cup \text{mayor2}) = P(\text{par}) + P(\text{mayor2}) - P(\text{par} \cap \text{mayor2}) = 3/6 + 4/6 - 2/6 = 5/6$ (observar que "par $\cap$ mayor2" = $\{4, 6\}$).

5. Distribuciones de probabilidad discretas

Al inicio de este módulo se definió el concepto de variable cuantitativa discreta como aquella variable cuantitativa que podía tomar un número finito o contable de valores distintos. Así, un ejemplo de variable discreta sería X = “resultado del lanzamiento de un dado”, ya que dicha variable sólo puede tomar seis posibles valores.

Cada uno de los posibles valores de una variable discreta tendrá asociada una probabilidad de ocurrencia (p. ej., en el caso del dado, la probabilidad de obtener un 2 será de 1/6), por lo que parece natural estudiar cómo se distribuyen o comportan dichas probabilidades. En concreto, se puede definir una “función de probabilidad”, $f(x)$, que asocie a cada valor x de la variable discreta X su probabilidad de ocurrencia, $P(x)$. Por ejemplo, en el caso de la variable anterior, asociada al experimento aleatorio **“lanzamiento de un dado normal”**, la correspondiente función de probabilidad sería: $f(1) = P(X = 1) = 1/6$, $f(2) = P(X = 2) = 1/6$, $f(3) = P(X = 3) = 1/6$, $f(4) = P(X = 4) = 1/6$, $f(5) = P(X = 5) = 1/6$, $f(6) = P(X = 6) = 1/6$.

Observad

Fijaos que si se usara un **dado “trucado”**, no todas las probabilidades de ocurrencia serían iguales y, por tanto, la función de probabilidad tomaría valores distintos para distintos valores posibles de la variable.

Dada una variable aleatoria discreta X, resulta útil conocer la **distribución de probabilidad** de dicha variable, es decir, cómo se distribuyen o comportan las probabilidades de ocurrencia de sus posibles valores. A tal efecto se definen las siguientes funciones:

La **función de probabilidad** de X es aquella función $f(x)$ que asigna a cada posible valor x de X su probabilidad de ocurrencia, es decir: $f(x) = P(X = x)$ para todo valor posible x de X.

La **función de distribución** de X es aquella función $F(x)$ que asigna a cada posible valor x de X su probabilidad acumulada de ocurrencia, es decir $F(x) = P(X \leq x)$ para todo valor posible x de X.

La tabla 3 muestra la función de probabilidad y la función de distribución correspondientes a la variable X anterior pero usando un dado “trucado” que tiene dos valores 6 y ningún valor 2. Por su parte, la figura 18 muestra ambas funciones superpuestas en el mismo gráfico. Observando detenidamente la tabla 3 y la figura 18 se pueden deducir las siguientes características propias de estas funciones:

- Puesto que representan probabilidades, ambas funciones siempre toman valores en el intervalo [0, 1].

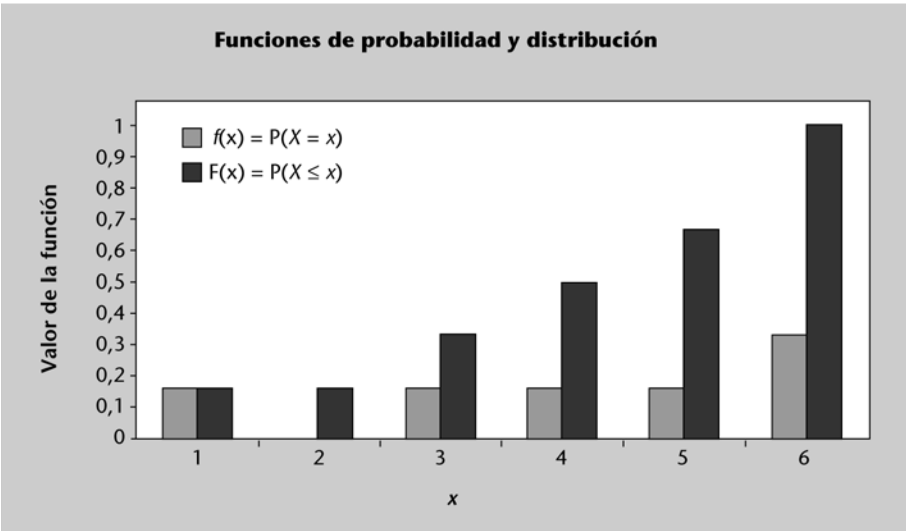
- La suma de todos los valores que toma la función de probabilidad siempre ha de ser 1 (ello se debe a las propiedades de la probabilidad).

La función de distribución siempre es una función creciente que pasa de valor 0 en su extremo izquierdo ($F(0) = P(X \leq 0) = 0$) a valor 1 en su extremo derecho ($F(6) = P(X \leq 6) = 1$).

Tabla 3. Funciones de probabilidad y distribución para una variable discreta

Variable X	Función de probabilidad $f(x) = P(X = x)$	Función de distribución $F(x) = P(X \leq x)$
1	1/6	1/6
2	0	1/6
3	1/6	2/6
4	1/6	3/6
5	1/6	4/6
6	2/6	1
Total	**1**	

Figura 18. Funciones de probabilidad y distribución de una variable discreta



Parámetros descriptivos de una distribución discreta

Mientras que los estadísticos descriptivos y los gráficos o tablas de frecuencias se utilizan para analizar el comportamiento (distribución) de una muestra de observaciones empíricas, las distribuciones de probabilidad son modelos estadísticos que usan parámetros y funciones de distribución para describir el comportamiento teórico (distribución teórica) de toda una población. De forma análoga a lo que ocurría con las muestras –que se caracterizan por estadísticos descriptivos como la media o la varianza muestral–, las distribuciones de probabilidad asociadas a poblaciones también suelen caracterizarse por parámetros tales como la media o la varianza poblacional. Ahora bien, puesto que en general no se dispondrá de observaciones sobre toda la población sino sólo de una función de distribución o de probabilidades, la forma de calcular dichos parámetros es algo distinta:

- **Media o valor esperado de una variable discreta:** la media o valor esperado de una variable discreta X que puede tomar los valores x_1, x_2, ..., se representa con μ o $E[X]$ y se calcula de la siguiente forma:

$$\mu = E[X] = x_1 \cdot P(X = x_1) + x_2 \cdot P(X = x_2) + ... = \sum_i x_i \cdot f(x_i)$$

donde $f(x)$ denota a la función de probabilidad de X.

Ejemplo: el caso de un dado equilibrado, el valor esperado o media de X = "resultado del lanzamiento" sería μ = 3; sin embargo, en el caso del dado "trucado" que se muestra en la tabla 3, la media o valor esperado es:

$$\mu = 1 \cdot f(1) + 2 \cdot f(2) + 3 \cdot f(3) + 4 \cdot f(4) + 5 \cdot f(5) + 6 \cdot f(6) =$$
$$= 1 \cdot \frac{1}{6} + 2 \cdot 0 + 3 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + 5 \cdot \frac{1}{6} + 6 \cdot \frac{2}{6} = 4{,}167$$

- **Varianza y desviación estándar de una variable discreta:** la varianza de una variable discreta X que puede tomar los valores x_1, x_2, ..., se representa con σ^2 y se calcula de la siguiente forma:

$$\sigma^2 = (x_1 - \mu)^2 \cdot P(X = x_1) + (x_2 - \mu)^2 \cdot P(X = x_2) + ... = \sum_i (x_i - \mu)^2 \cdot f(x_i)$$

donde $f(x)$ denota a la función de probabilidad de X. De forma análoga a cómo ocurría con los estadísticos muestrales, la desviación estándar de una variable es la raíz cuadrada positiva de su varianza, es decir:

$$\sigma = \sqrt{\sigma^2}$$

Ejemplo: en el caso del dado "trucado" que se muestra en la tabla 3, la varianza es:

$$\sigma^2 = (1 - 4{,}167)^2 \cdot \frac{1}{6} + (2 - 4{,}167)^2 \cdot 0 + (3 - 4{,}167)^2 \cdot \frac{1}{6} +$$
$$+ (4 - 4{,}167)^2 \cdot \frac{1}{6} + (5 - 4{,}167)^2 \cdot \frac{1}{6} + (6 - 4{,}167)^2 \cdot \frac{2}{6} = 3{,}139$$

Y la correspondiente desviación estándar: $\sigma = \sqrt{3{,}139} = 1{,}772$

La distribución binomial

Una de las distribuciones discretas más usadas en la práctica es la distribución binomial. Esta distribución se usa para contestar a preguntas como las siguientes:

- Si cada vez que un sistema informático es atacado por un virus la probabilidad de que el sistema no falle es de 0,76, ¿cuál es la probabilidad de que no se haya producido ningún fallo en el sistema tras cinco ataques?

- Si cada vez que se consulta una fuente de información la probabilidad de que ésta proporcione una respuesta satisfactoria es de 0,85, ¿cuál es la probabilidad de que se obtenga alguna respuesta satisfactoria tras tres consultas?

- Si tras la administración de un fármaco a un paciente en estado crítico la probabilidad de supervivencia de éste es de 0,99, ¿cuál es la probabilidad de que sobrevivan los catorce pacientes críticos que han recibido el tratamiento?

- Si la probabilidad de obtener una concesión para un proyecto de investigación es de 0,20, ¿cuál es la probabilidad de obtener al menos una concesión tras tres intentos?

- Si cada vez que se trata de encuestar a un transeúnte elegido al azar la probabilidad de que responda es de 0,15, ¿cuál es la probabilidad de que se consigan obtener ochenta respuestas o más a partir de una muestra aleatoria de ciento cincuenta transeúntes?

Distribución de Poisson y la uniforme discreta

Otras distribuciones discretas muy habituales son la distribución de Poisson y la uniforme discreta. Es posible encontrar en Internet abundante documentación sobre éstas y otras distribuciones discretas así como sobre sus ámbitos de aplicación.

La **distribución binomial** es un modelo estadístico que permite calcular probabilidades sobre la variable aleatoria X = "número de éxitos conseguidos en n pruebas independientes". Cada una de estas n pruebas es una repetición de un experimento aleatorio cuyo resultado es binario (éxito o fracaso), siendo p la probabilidad de "éxito" en cada prueba y $q = 1 - p$ la probabilidad de "fracaso".

Resultado "éxito"

No debe confundirse el resultado **"éxito"** de un experimento aleatorio con el hecho de que el resultado sea deseable desde un punto de vista social o subjetivo. Así, por ejemplo, se podría considerar "éxito" del experimento aleatorio el fallo del sistema informático que sufre el ataque de un virus.

Cabe observar que la variable X = "número de éxitos en n pruebas independientes" puede tomar cualquier valor k entre 0 y n (ambos inclusive). Se suele usar la notación $X \sim B(n, p)$ para indicar que X se distribuye o se comporta según una distribución binomial de parámetros n (número de pruebas o repeticiones) y p (probabilidad de "éxito" en cada prueba). En tales condiciones, las probabilidades asociadas a dicha variable vienen dadas por la expresión matemática siguiente:

Para cualquier k entre 0 y n, $P(X = k) = \binom{n}{k} p^k \cdot (1-p)^{n-k}$, donde $\binom{n}{k} = \frac{n!}{k!(n-k)!}$,

siendo $0! = 1! = 1$ y $n! = n \cdot (n-1) \dots 1$ para todo $n > 1$.

Observad

La expresión "$n!$" se lee como "factorial de n" o "n factorial". Así, por ejemplo, $4! = 4 \cdot 3 \cdot 2 \cdot 1$ y $6! = 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1$. Sin embargo, $1! = 1$ y $0! = 1$.

Se cumple, además, que la media (valor esperado) y la varianza de una distribución binomial son, respectivamente: $\mu = n \cdot p$ y $\sigma^2 = n \cdot p \cdot (1 - p)$.

Ejemplo: la probabilidad de que al introducir datos en un formulario web se cometa un error es de 0,1. Si diez personas rellenan el formulario de forma independiente, ¿cuál es la probabilidad de que no haya más de un formulario erróneo?, ¿cuál es el valor esperado y la desviación estándar de la variable considerada?

Fijémonos en que, en este caso, X = "número de formularios erróneos en diez pruebas" y $X \sim B$ (10, 0,1). Además, se pide $P(X \le 1) = P(X = 0 \cup X = 1) = P(X = 0) + P(X = 1)$ (puesto que son sucesos disjuntos). Ahora bien:

$$P(X=0)=\binom{10}{0}0{,}1^{0}\cdot(0{,}9)^{10}=\frac{10!}{0!10!}(1)(0{,}3487)=0{,}3487$$

$$P(X=1)=\binom{10}{1}0{,}1^{1}\cdot(0{,}9)^{9}=\frac{10!}{1!9!}(0{,}1)(0{,}3874)=0{,}3874$$

Por tanto, $P(X \le 1) = 0{,}3874 + 0{,}3487 = 0{,}7361$. Finalmente, $\mu = 10 \cdot 0{,}1 = 1$ y $\sigma = \sqrt{10 \cdot 0{,}1 \cdot 0{,}9} = 0{,}9487$.

En la práctica, los cálculos probabilísticos anteriores se suelen automatizar con la ayuda de algún programa estadístico o de análisis de datos. La figura 19 muestra cómo se pueden calcular probabilidades de una binomial con ayuda de Minitab. La figura 20, por su parte, muestra cómo obtenerlas usando Excel.

Pasos a seguir

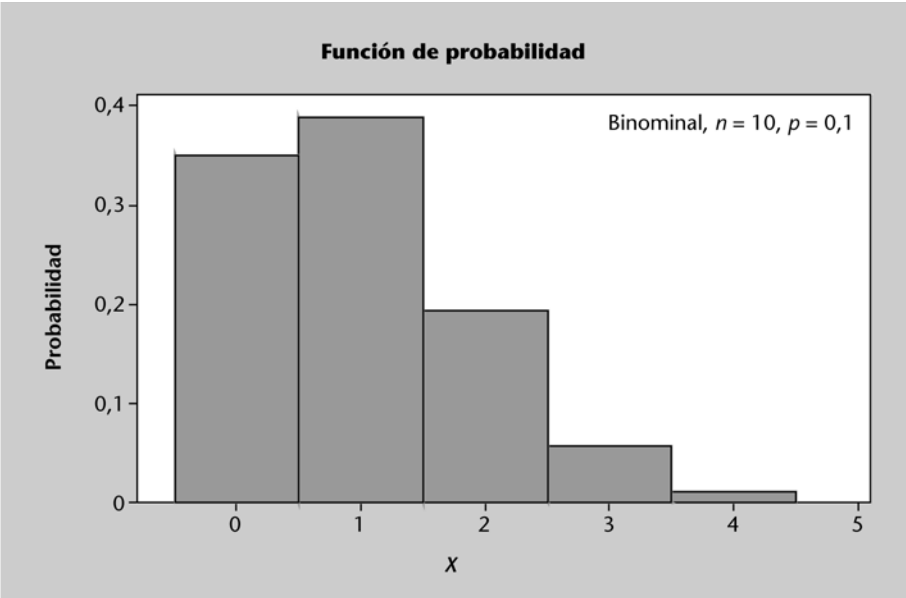
Se sigue la ruta ***Calc*** > ***Probability Distributions*** > ***Binomial (1)*** y se completan los parámetros en la ventana correspondiente ***(2)***. El resultado se muestra en ***(3)***. Observar que, si en lugar de escoger la opción ***Cumulative probability*** en ***(2)*** se hubiera escogido la opción ***Probability***, el programa hubiera calculado $P(X = 1)$ en lugar de $P(X <= 1)$. Finalmente, para una probabilidad p dada, la opción ***Inverse cumulative probability*** devuelve aquel valor c de la variable X tal que $P(X <= c) = p$.

Figura 19. Cálculo de probabilidades en una binomial con Minitab y Excel



La figura 20 se muestra la función de probabilidad asociada a la binomial del ejemplo anterior. Se observa que, aunque en teoría los posibles valores de la variable *X* irían desde 0 hasta 10 (número de pruebas), en la práctica los valores mayores de 4 tienen probabilidad de suceso prácticamente nula (por ejemplo, es muy poco frecuente que se obtengan valores superiores a 4). En efecto, $P(X > 4) = 1 - P(X <= 4)$ = {usando Minitab o Excel} = 1 – 0,9984 = 0,0016.

Figura 20. Función de probabilidad de una *B* (10, 0,1)



Las probabilidades anteriores se pueden obtener también mediante el uso de tablas estadísticas (sin necesidad de usar ningún software). Así, siguiendo el ejemplo anterior, la figura 21 muestra cómo calcular $P(X = 1)$ usando la tabla binomial. En este caso, *X* es una $B(10, 0{,}1)$ y se quiere hallar $P(X = k)$ siendo $k = 1$. Para ello, se busca la sección de la tabla correspondiente a $n = 10$, y la intersección entre la fila $k = 1$ y la columna $p = 0{,}1$.

Cálculo de probabilidades

Resulta fácil encontrar en Internet abundantes documentos que explican con todo detalle el uso de tablas para calcular probabilidades. En la medida de lo posible, sin embargo, conviene automatizar los cálculos mediante el uso de software.

Figura 21. Cálculo de probabilidades binomiales mediante tablas

n	*k*	*p*	0,01	0,05	0,10	0,15	0,20	0,25
	7		0,0000	0,0000	0,0000	0,0000	0,0001	0,0004
	8		0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
9	0		0,9135	0,6302	0,3874	0,2316	0,1342	0,0751
	1		0,0830	0,2985	0,3874	0,3679	0,3020	0,2253
	2		0,0034	0,0629	0,0446	0,2597	0,3020	0,3003
	3		0,0001	0,0077	0,0074	0,1069	0,1762	0,2336
	4		0,0000	0,0006	0,0008	0,0283	0,0661	0,1168
	5		0,0000	0,0000	0,0001	0,0050	0,0165	0,0389
	6		0,0000	0,0000	0,0000	0,0006	0,0028	0,0087
	7		0,0000	0,0000	0,0000	0,0000	0,0003	0,0012
	8		0,0000	0,0000	0,0000	0,0000	0,0000	0,0001
	9		0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
10	0		0,9044	0,5987	0,3487	0,1969	0,1074	0,0563
	1		0,0914	0,3151	0,3874	0,3474	0,2684	0,1877
	2		0,0042	0,0746	0,1937	0,2759	0,3020	0,2816
	3		0,0001	0,0105	0,0574	0,1298	0,2013	0,2503
	4		0,0000	0,0010	0,0112	0,0401	0,0881	0,1460
	5		0,0000	0,0001	0,0015	0,0085	0,0264	0,0584

p

$P(X = 1) = 0{,}3874$

n

k

6. Distribuciones de probabilidad continuas

Al inicio de este módulo se definió el concepto de variable cuantitativa continua como aquella variable cuantitativa que podía tomar un número infinito (no contable) de valores distintos. Así, un ejemplo de variable continua sería X = “tiempo que se tarda en desarrollar un portal web”, ya que esta variable puede tomar un valor real cualquiera entre 0 e infinito.

A diferencia de lo que ocurría con las variables discretas, cuando se trabaja con variables continuas no es posible definir una función de probabilidad que asigne probabilidades a los distintos valores de la variable: si X es una variable continua, X puede tomar un número infinito (no contable) de valores, por lo que la probabilidad teórica de que la variable X tome un valor concreto x es siempre 0, es decir: $P(X = x) = 0$ para cualquier valor x de X. Sí es posible, sin embargo, asignar probabilidades a intervalos de valores. Por ejemplo, si el 51% de los portales web tardan en desarrollarse entre 240 y 258 horas, entonces $P(240 < X < 258) = 0{,}51$. Para describir la distribución de probabilidad de una variable continua se sigue usando la función de distribución (aunque con algún matiz nuevo) y, además, se usa también la llamada “función de densidad” en lugar de la función de probabilidad típica de variables discretas:

Nota

En variables continuas, puesto que $P(X = x) = 0$ para cualquier valor x de X, se cumplirá que:

a) $P(X \leq x) = P(X < x)$

b) $P(X \geq x) = P(X > x)$

La **función de densidad** de una variable continua X es una función $f(x)$ tal que la probabilidad de que X tome un valor en un intervalo (a, b) coincide con el **área** “encerrada” por dicha función entre los extremos de dicho intervalo (figura 22), es decir: $P(a < X < b)$ = área bajo $f(x)$ entre a y b.

La **función de distribución** de X es aquella función $F(x)$ que asigna a cada posible valor x de X su probabilidad acumulada de ocurrencia (figura 23), es decir, $F(x) = P(X \leq x)$ = área bajo $f(x)$ desde $-\infty$ (menos infinito) hasta x.

Nota

La función de densidad $f(x)$ siempre es positiva y “encierra” un área total de 1.

Atención

Observar la equivalencia entre los conceptos de “probabilidad” y “área”.

La figura 22 muestra la función de densidad de una variable con distribución simétrica y centrada en el valor 250 (puesto que la función es totalmente simétrica la media y la mediana coinciden en este punto). Se observa también el área encerrada bajo función de densidad entre los valores $a = 240$ y $b = 258$. Esta área corresponde con la probabilidad siguiente: $P(240 < X < 258)$. Por su parte, la figura 23 muestra la función de distribución asociada a la misma variable. Nuevamente se aprecia la simetría con respecto al valor central, así como el hecho de que la función de distribución va creciendo conforme va acumulando probabilidades, pasando del valor 0 en su extremo izquierdo al valor 1 en su extremo derecho. A partir de esta gráfica se pueden estimar visualmente probabilidades acumuladas, por ejemplo: $P(X <= 260)$ será un valor muy cercano a 0,8.

Figura 22. Función de densidad de una variable continua y área encerrada

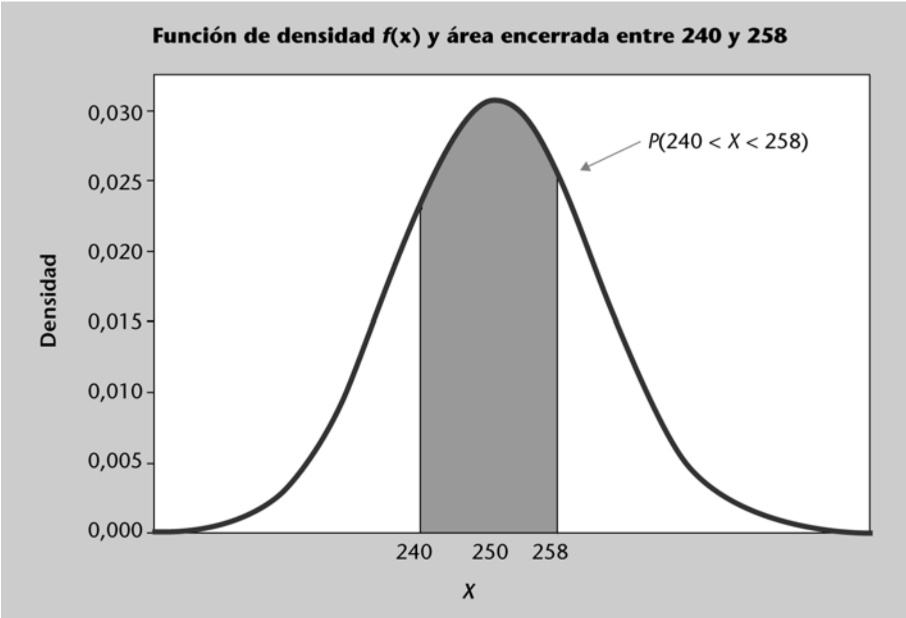
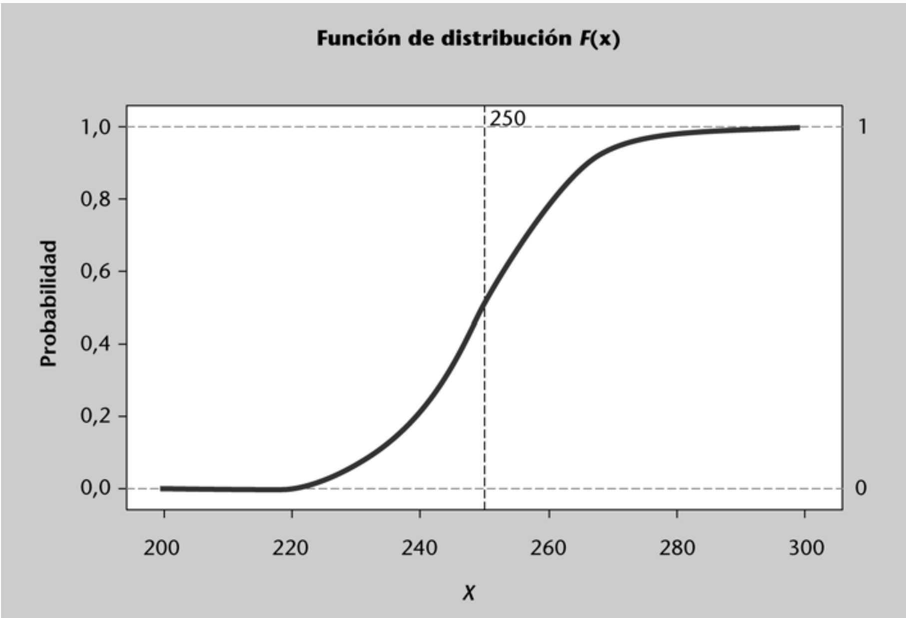


Figura 23. Función de distribución de una variable continua



Función de distribución

La función de distribución es una función acumulativa de probabilidades y, por tanto, es siempre creciente, pasando de 0 (extremo izquierdo) a 1 (extremo derecho).

Parámetros descriptivos de una distribución continua

En el caso de distribuciones continuas, la forma de calcular los parámetros es similar a la empleada para distribuciones discretas, si bien ahora los sumatorios se sustituyen por áreas (integrales definidas en términos matemáticos) entre dos extremos:

Atención

Aunque en la práctica se hará uso de programas estadísticos para hacer los cálculos, es importante conocer qué conceptos se usan para definir cada tipo de parámetro.

- **Media o valor esperado de una variable continua:** la media o valor esperado de una variable continua X se representa por μ o $E[X]$ y se calcula de la siguiente forma:

$$\mu = E[X] = \text{área total bajo “}x \cdot f(x)\text{”} = \int_{-\infty}^{+\infty} x \cdot f(x)dx$$

donde $f(x)$ denota a la función de densidad de X.

- **Varianza y desviación estándar de una variable continua:** la varianza de una variable continua X se representa por σ^2 y se calcula de la siguiente forma:

$$\sigma^2 = \text{área total bajo “}(x-\mu)^2 \cdot f(x)\text{”} = \int_{-\infty}^{+\infty} (x-\mu)^2 \cdot f(x)\,dx$$

  donde $f(x)$ denota a la función de densidad de X. Como siempre, la desviación estándar de una variable es la raíz cuadrada positiva de su varianza, es decir:

$$\sigma = \sqrt{\sigma^2}$$

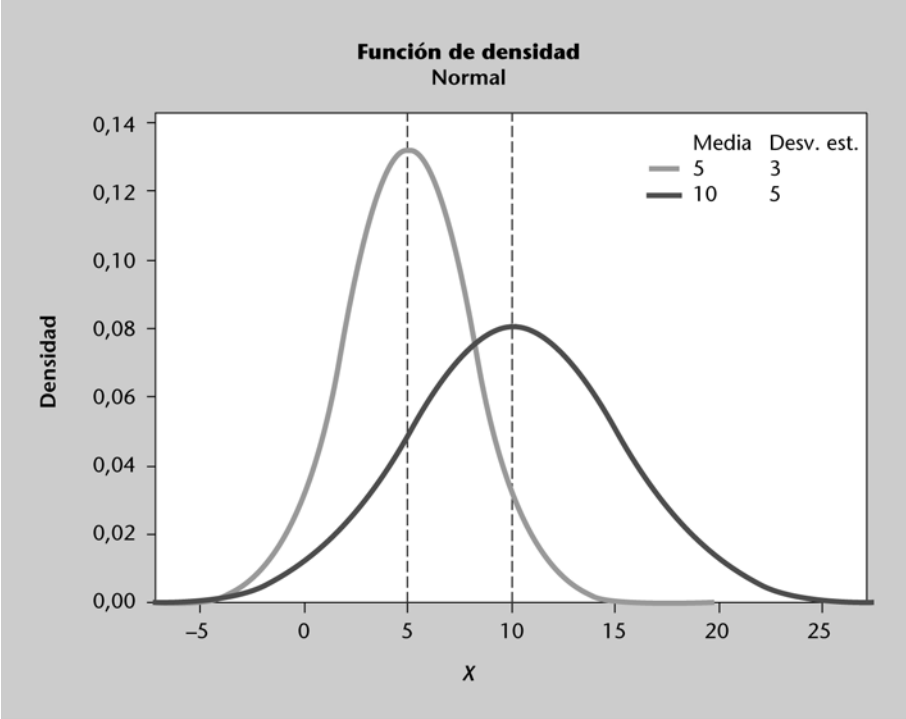
La distribución normal o gaussiana

La distribución normal o gaussiana es la distribución teórica más importante. Muchas variables continuas siguen una distribución normal o aproximadamente normal. Otras variables continuas y discretas también pueden, en determinadas circunstancias, ser aproximadas mediante una distribución normal. La normal, además, es una distribución clave en la estadística inferencial ya que algunas de sus propiedades se utilizan para obtener información sobre toda la población a partir de información sobre una muestra.

La forma concreta de una distribución normal viene caracterizada por dos parámetros: la media, μ, que define dónde se sitúa el centro de la función de densidad, y la desviación estándar, σ, que define la amplitud de la función de densidad. Cuando una variable continua X sigue una distribución normal, se suele representar por $X \sim N(\mu, \sigma)$.

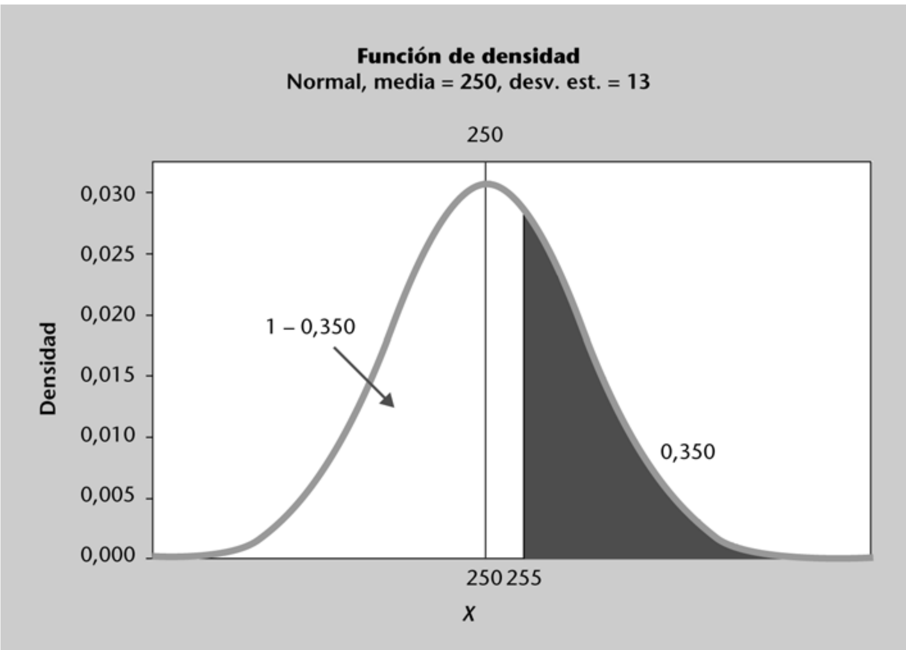
Las figuras 22 y 23 muestran, respectivamente, la función de densidad y la función de distribución de una normal con media $\mu = 250$ y desviación estándar $\sigma = 13$. La figura 24 muestra las funciones de densidad para dos distribuciones de tipo normal con parámetros $\{\mu = 5, \sigma = 3\}$ y $\{\mu = 10, \sigma = 5\}$ respectivamente. Se observa que la función de densidad de la normal tiene forma de “campana de Gauss”, elevada en el centro (el valor medio o esperado) y con dos colas simétricas en los extremos. Es de destacar, además, cómo cada una de las curvas está centrada en su media, así como el hecho de que la curva es más ancha cuanto mayor es la desviación estándar.

Figura 24. Funciones de densidad asociadas a sendas normales



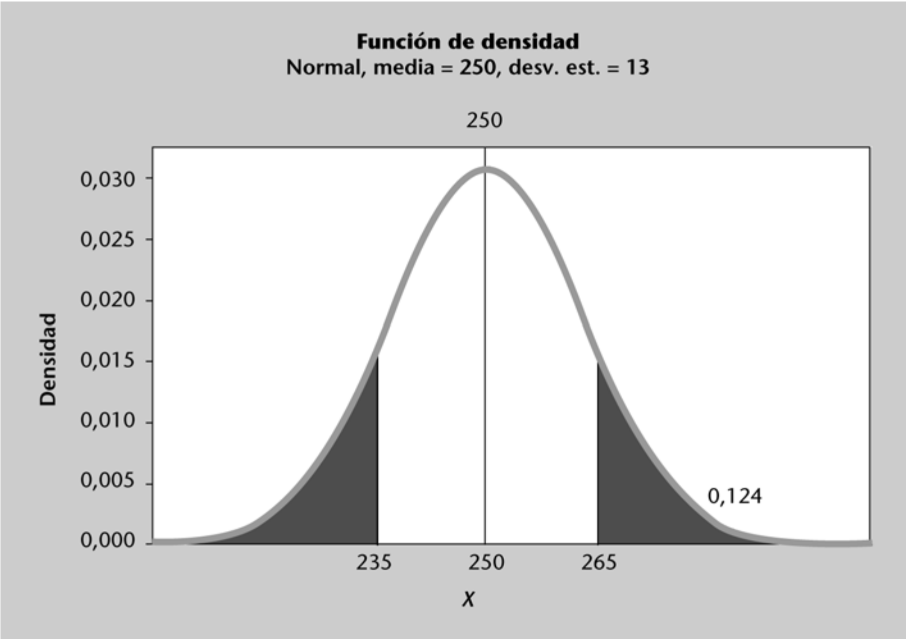
Como en cualquier otra función de densidad, el área total encerrada bajo la curva es de 1. En la práctica eso significa que para cualquier valor x de X, $P(X > x) = 1 - P(X < x)$, es decir, el área a la derecha de un valor es el área total (que vale 1) menos el área a su izquierda y viceversa (figura 25). Además, puesto que la normal es una distribución simétrica con respecto a su media, el área "encerrada" por una cola es igual al área "encerrada" por la cola opuesta (figura 26).

Figura 25. El área total de una función de densidad es 1



Cualquier distribución normal cumple además la llamada **regla 68-95-99,7** según la cual el intervalo (μ – σ, μ + σ) contiene aproximadamente el 68% de las observaciones, el intervalo (μ – 2σ, μ + 2σ) contiene aproximadamente el 95% de las observaciones y el intervalo (μ – 3σ, μ + 3σ) contiene aproximadamente el 99,7% de las observaciones. Así, por ejemplo, si $X \sim N$ (250,13) se puede afirmar que un 68% de las observaciones de *X* estarán en el intervalo (237, 263), un 95% de las observaciones estarán en el intervalo (224, 276) y un 99,7% de las observaciones estarán en el intervalo (211, 289). Observad, por tanto, que será altamente improbable encontrar valores de *X* fuera de este último intervalo.

Figura 26. Dos colas simétricas "encierran" la misma área



De entre las infinitas distribuciones normales que se pueden considerar variando los parámetros μ y σ conviene citar la llamada **normal estándar**, que tiene por parámetros μ = 0 y σ = 1. En otras palabras, una variable continua *Z* se distribuirá según una normal estándar, $Z \sim N$ (0,1), si su función de densidad es la de una normal centrada en el origen y con desviación estándar unitaria. Esta distribución normal estándar se suele usar bastante en estadística inferencial y también cuando se desean calcular probabilidades de una normal cualquiera mediante el uso de tablas de probabilidades ya calculadas.

En efecto, dada una variable normal cualquiera, $X \sim N$ (μ, σ), es posible aplicarle un **proceso de estandarización** para obtener una normal estándar *Z*. Esto se consigue restando a la variable *X* su media μ (con lo que la función de densidad se desplaza a lo largo del eje *x* hasta que queda centrada en el origen) y dividiendo el resultado por su desviación estándar σ (con lo que la nueva variable tendrá una desviación estándar unitaria), es decir: $Z = \frac{X - \mu}{\sigma} \sim N(0,1)$. Este proceso de estandarización permite, entre otras cosas, calcular probabilidades para una normal cualquiera a partir de las tablas de probabilidades precalculadas que existen para la distribución

normal estándar, lo que evita el tener que resolver integrales cada vez que se desea obtener una nueva probabilidad. Supongamos, por ejemplo, que X sigue una $N(1.500, 100)$ y se desea obtener $P(X < 1.400)$ mediante el uso de tablas. El primer paso consiste en estandarizar los valores:

$$P(X < 1.400) = P\left(\frac{X-\overline{x}}{\sigma} < \frac{1.400-\overline{x}}{\sigma}\right) = P\left(Z < \frac{1.400-1.500}{100}\right) = P(Z < -1)$$

En otras palabras, se desea calcular el área a la izquierda del valor –1 en una normal tipificada o estándar. Normalmente, la tabla de la normal estándar, Z, ofrece áreas (probabilidades) a la izquierda de valores positivos, por lo que resultará necesario hacer una pequeña transformación teniendo en cuenta que: (a) por simetría de la normal estándar, el área (probabilidad) a la izquierda de un valor negativo k es igual al área (probabilidad) a la derecha del correspondiente valor positivo, $|k|$ (p. ej., $P(Z < -1) = P(Z > 1)$), y (b) el área (probabilidad) total encerrada bajo la curva es 1 (p. ej., el área a la izquierda de un valor más el área a su derecha suma 1, por ejemplo: $P(Z < 1) + P(Z > 1) = 1$). Teniendo en cuenta lo anterior, se deduce que $P(Z < -1) = P(Z > 1) = 1 - P(Z < 1)$ = {ver tabla figura 27} = 1 – 0,8413 = 0,1587.

Figura 27. Cálculo de probabilidades en una normal mediante tablas

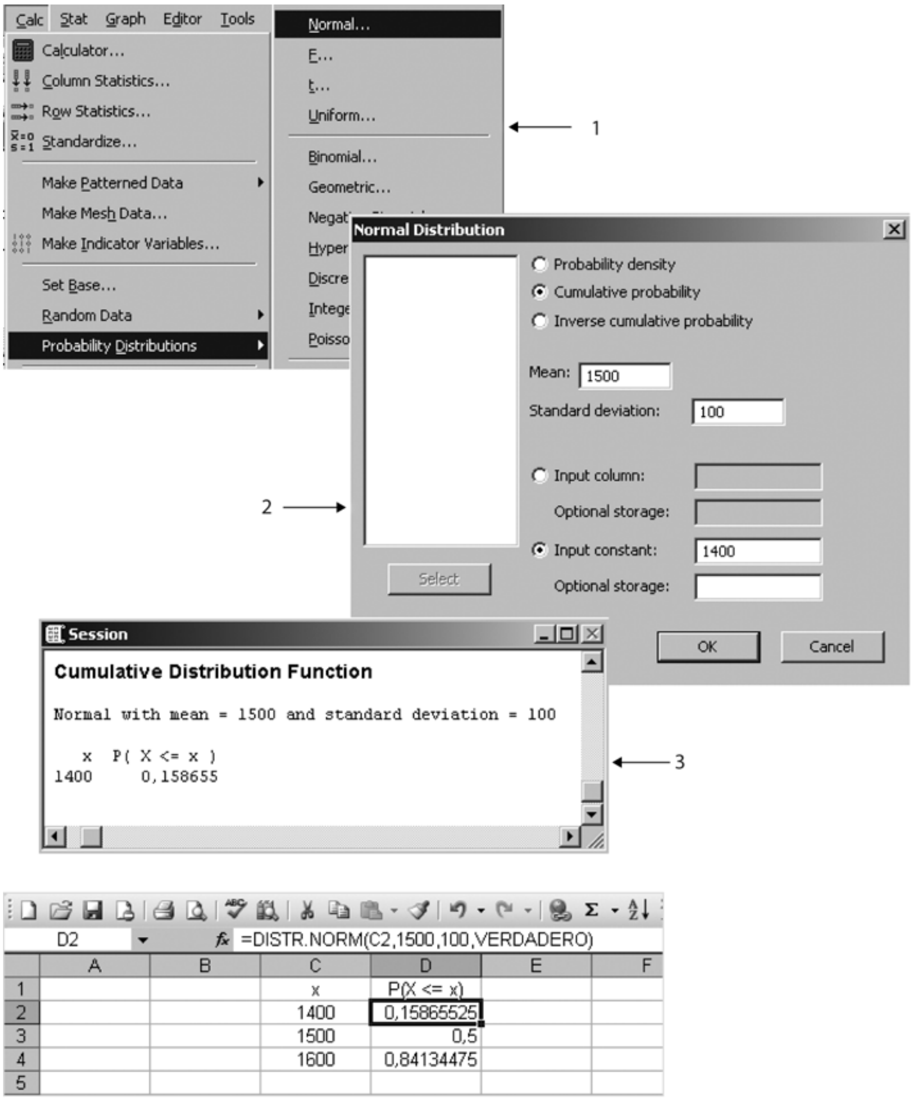
	,00	,01	,02	,03	,04	,05
0,5	0,6915	0,6950	0,6985	0,7019	0,7054	0,7088
0,6	0,7257	0,7291	0,7324	0,7357	0,7389	0,7422
0,7	0,7580	0,7611	0,7642	0,7673	0,7704	0,7734
0,8	0,7881	0,7910	0,7939	0,7967	0,7995	0,8023
0,9	0,8159	0,8186	0,8212	0,8238	0,8264	0,8289
1,0	0,8413	0,8438	0,8461	0,8485	0,8508	0,8531
1,1	0,8643	0,8665	0,8686	0,8708	0,8729	0,8749
1,2	0,8849	0,8869	0,8888	0,8907	0,8925	0,8944
1,3	0,9032	0,9049	0,9066	0,9082	0,9099	0,9115
1,4	0,9192	0,9207	0,9222	0,9236	0,9251	0,9265

Nota

Notar que para hallar $P(Z < 1{,}00)$ usando la tabla se ha de buscar el valor intersección entre la fila 1,0 y la columna 0,00 (dado que 1,00 = 1,0 + 0,00). Si se pidiese $P(Z < 1{,}24)$, entonces habría que buscar la intersección entre la fila 1,2 y la columna 0,04 (dado que 1,24 = 1,2 + 0,04), con lo que se obtendría el valor 0,8925.

Por otra parte, también es posible automatizar el cálculo de probabilidades de una normal cualquiera mediante el uso de programas estadísticos, con lo que se elimina así la necesidad de resolver manualmente las integrales indefinidas o de tener que usar tablas de probabilidades precalculadas. La figura 28 muestra cómo obtener probabilidades de una normal con Minitab. En concreto, para una normal con media $\mu = 1.500$ y desviación estándar $\sigma = 100$, se obtiene que $P(X < 1.400) = 0{,}158655$. Asimismo, la figura 28 muestra cómo se han obtenido con Minitab y Excel algunas probabilidades para la misma variable. Es preciso observar que $P(X < 1.500) = 0{,}5$, lo cual es lógico puesto que 1.500 es la media y, a la vez, la mediana de la distribución normal.

Figura 28. Cálculo de probabilidades en una normal con Minitab y Excel



Pasos a seguir

Se sigue la ruta ***Calc*** > ***Probability Distributions*** > ***Normal (1)*** y se completan los parámetros en la ventana correspondiente ***(2)***. El resultado se muestra en ***(3)***. Observar que, si en lugar de escoger la opción ***Cumulative probability*** en ***(2)*** se hubiera escogido la opción ***Probability density***, el programa hubiera calculado el valor de la función de densidad en $x = 1.400$ en lugar de $P(X < 1.400)$. Finalmente, para una probabilidad p dada, la opción ***Inverse cumulative probability*** devuelve aquel valor c de la variable X tal que $P(X < c) = p$.

Ejemplos de aplicación de una normal

- Según un estudio realizado por el Ministerio de Educación, el número de horas anuales que dedican los niños españoles a ver la televisión es una variable aleatoria que sigue una distribución normal de media 1.500 horas y desviación estándar de 100 horas. ¿Qué porcentaje de niños dedican entre 1.400 y 1.600 horas anuales?

  En este caso, $X \sim N$ (1.500,100) y se pide $P(1.400 < X < 1.600)$. Por la regla 68-95-99,7, se tiene que la probabilidad anterior será, aproximadamente, del 68% (ya que $\mu - \sigma = 1.400$ y $\mu + \sigma = 1.600$). Para calcular de forma más exacta dicha probabilidad, conviene notar que $P(1.400 < X < 1.600) = P(X < 1.600) - P(X < 1.400)$, es decir: el área entre 1.400 y 1.600 coincide con el área a la izquierda de 1.600 menos el área a la izquierda de 1.400. Las probabilidades anteriores se pueden calcular usando cualquier programa estadístico (p. ej.: Minitab o Excel), y resultan: $P(X < 1.600) = 0{,}8413$ y $P(X < 1.400) = 0{,}1587$, por lo que la probabilidad buscada es de 0,6827, es decir, un 68,27% de los niños dedican entre 1.400 y 1.600 horas anuales a ver la televisión.

- En base a los datos del Instituto Nacional de Estadística (INE), el sueldo medio anual de un trabajador es de 26.362 euros. Suponiendo que dichos sueldos sigan una distribución normal con una desviación estándar de 6.500 euros, ¿cuál será el porcentaje de trabajadores que superen los 40.000 euros?

  En este caso, $X \sim N$ (26.362,6.500) y se pide $P(X > 40.000)$. Observar que, puesto que el área total bajo la curva normal es 1, $P(X > 40.000) = 1 - P(X < 40.000)$ = {Minitab o Excel} = 1 – 0,9821 = 0,0179, es decir, sólo un 1,8% de los trabajadores superarían la cifra de los 40.000 euros anuales.

- El tiempo que se emplea en rellenar un cuestionario en línea sigue una distribución aproximadamente normal con una media de 3,7 minutos y una desviación estándar de 1,4 minutos. ¿Cuál es la probabilidad de que se tarde menos de 2 minutos en responder a dicho cuestionario? ¿Y de que se tarde más de 6 minutos? Hallad el valor c tal que $P(X < c) = 0,75$ (percentil 75 de la variable).

  En este caso, $X \sim N$ (3,7, 1,4). En primer lugar, $P(X < 2)$ = {Minitab o Excel} = 0,1131, es decir: un 11,31% de los individuos que respondan el cuestionario emplearan menos de 2 minutos en hacerlo. Por otra parte, $P(X > 6) = 1 - P(X < 6)$ = {Minitab o Excel} = 0,0505, es decir, un 5% de los individuos tardarán más de 6 minutos en responder el cuestionario. Finalmente, para hallar el valor c tal que $P(X < c) = 0,75$ se debe usar la opción *Inverse cumulative probability* de Minitab (o su equivalente en Excel), con lo que se obtiene un valor aproximado de 4,64 minutos, es decir el 75% de los individuos tardan menos de 4,64 minutos en completar el cuestionario (o, dicho de otro modo, el 25% tardan más de 4,64 minutos en hacerlo).

Las distribuciones *t*-Student y *F*-Snedecor

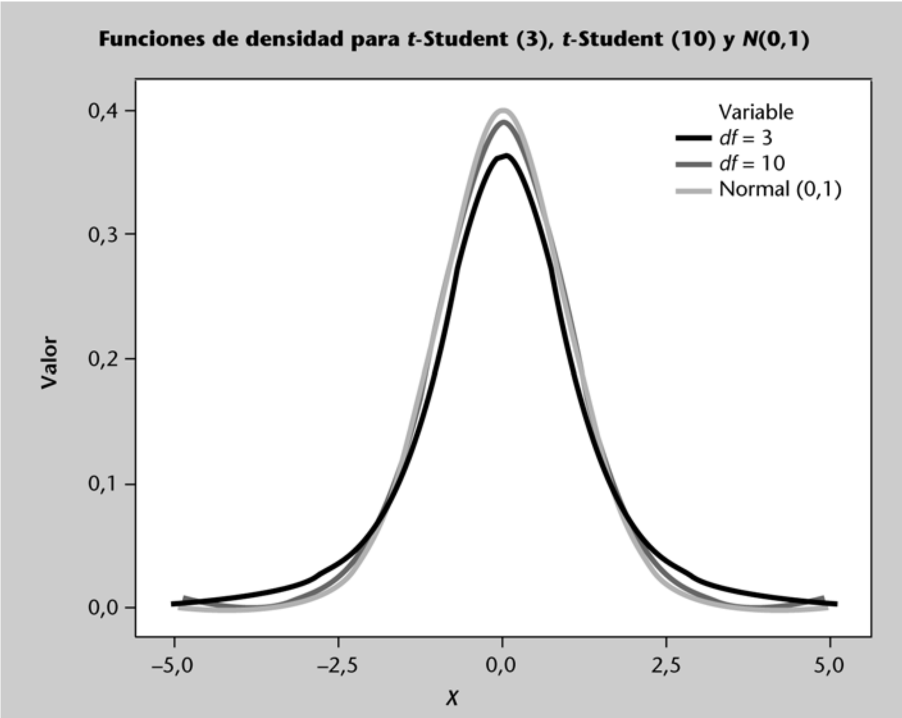
Además de la normal, hay muchas otras distribuciones de probabilidad continuas que se suelen usar en estadística inferencial. Una de ellas es la llamada distribución *t*-Student, y otra es la llamada *F*-Snedecor. Ambas se presentan a continuación:

La distribución ***t*-Student** es una distribución simétrica y centrada en el origen (es decir, su media y su mediana son 0). Esta distribución se caracteriza por un parámetro llamado **grados de libertad** o ***df*** (*degrees of freedom*), siendo $df > 2$. En la práctica, $df = n - 1$, donde n es el tamaño de la muestra que se esté analizando. La figura 29 muestra diversas funciones de densidad de las *t*-Student, cada una de ellas asociadas a un valor concreto del parámetro *df*. Se observa cómo la *t*-Student se asemeja cada vez más a una normal estándar conforme se va incrementando el parámetro grados de libertad.

Grados de libertad

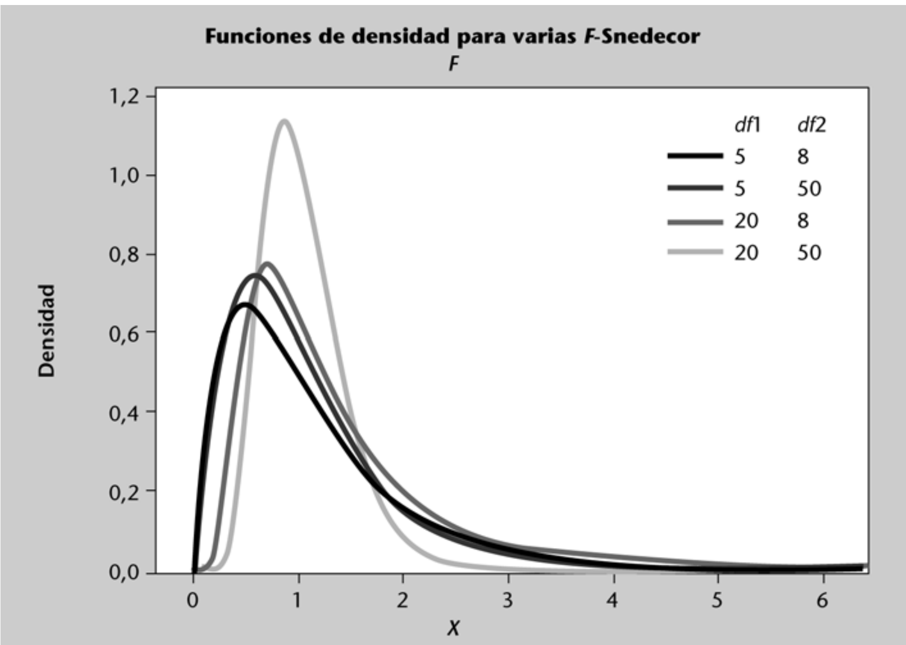
En estadística, el concepto de **grados de libertad** asociados a un conjunto de datos se puede interpretar como el número mínimo de valores que se necesitaría conocer para determinar dichos datos. Así, por ejemplo, en el caso de un muestra aleatoria de tamaño N, habría N grados de libertad (no se puede determinar el valor de ninguno de los datos incluso aunque se conociese el valor de los $N - 1$ restantes). Sin embargo, un conjunto de N datos de los cuales se conozcan $N - 1$, la media muestral tendría $N - 1$ grados de libertad (fijados los valores de los $N - 1$ datos y de la media, quedaría ya fijado el valor desconocido restante). Así, si tenemos un conjunto de 3 observaciones de la variable X, $x_1 = 2$, $x_2 = -2$ y $x_3 = a$ (desconocido), y sabemos que la media de los tres valores es 0, necesariamente $a = 0$.

Figura 29. Funciones de densidad de *t*-Student según *df*



Por su parte, la distribución ***F*-Snedecor** es otra distribución continua. La *F*-Snedecor siempre toma valores no negativos (es decir, una variable que siga dicha distribución sólo puede tomar valores iguales o mayores a 0, nunca valores negativos). Además, esta distribución no es simétrica, sino que está sesgada a la derecha (figura 30). Así como la normal venía caracterizada por dos parámetros, μ (media) y σ (desviación estándar), la *F*-Snedecor también se caracteriza por dos parámetros: los **grados de libertad del numerador**, *df*1 y los **grados de libertad del denominador**, *df*2. Al igual que ocurría con la *t*-Student, para cada valor de estos parámetros se obtiene una función de densidad distinta y, por tanto, una distribución *F*-Snedecor distinta.

Figura 30. Funciones de densidad de *t*-Student según *df*1 y *df*2

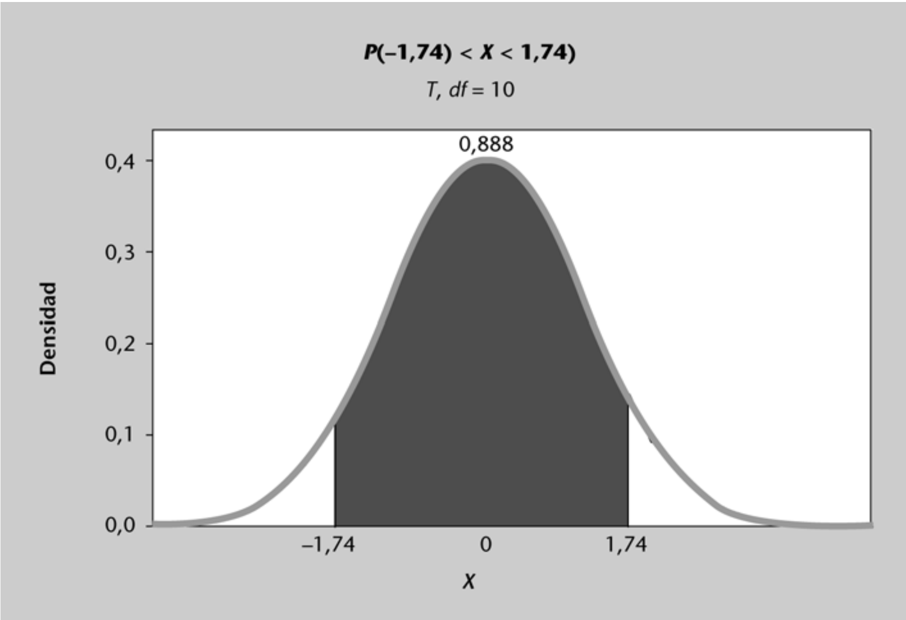


Para calcular probabilidades asociadas a una *t*-Student o a una *F*-Snedecor, pueden usarse programas estadísticos o de análisis de datos (Minitab, Excel, etc.) de forma análoga a como se hacía en el caso de la normal. Así, por ejemplo, si X es una variable aleatoria que sigue una distribución *t*-Student con diez grados de libertad, $P(-1{,}74 < X < 1{,}74) = P(X < 1{,}74) - P(X < -1{,}74)$ = {Minitab o Excel} = 0,9438 – 0,0562 = 0,8876 (figura 31).

Nota

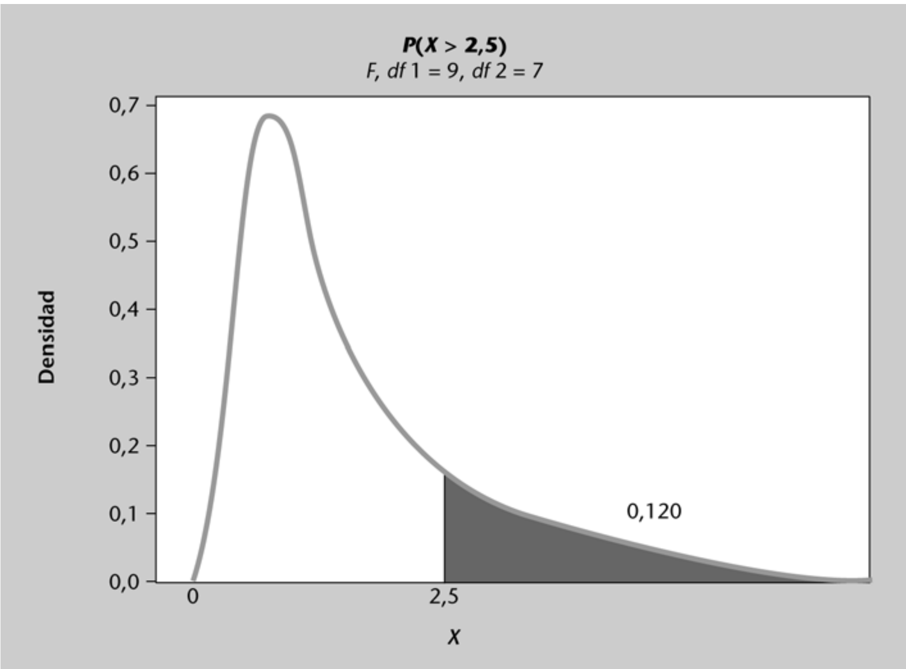
Notar que $P(-1{,}74 < X < 1{,}74)$ viene representada por el área marcada en la figura 31 (esto es, el área comprendida entre los valores –1,74 y 1,74). Para calcular dicha área, se calcula $P(X < 1{,}74)$ (p. ej., el área a la izquierda del 1,74) y al valor obtenido se le resta $P(X < -1{,}74)$ (p. ej., el área a la izquierda del -1,74). Para calcular $P(X < 1{,}74)$ con Minitab se usa el menú ***Calc*** > ***Probability Distributions*** > ***t…***, especificando los grados de libertad (10 en este ejemplo) y el valor de la constante (1,74 en este caso). Análogamente se obtendría el valor de $P(X < -1{,}74)$.

Figura 31. Probabilidades en una *t*-Student



Finalmente, si X es una variable aleatoria que sigue una distribución *F*-Snedecor con nueve grados de libertad en el numerador y siete grados de libertad en el denominador, entonces $P(X > 2{,}5) = 1 - P(X < 2{,}5)$ = {Minitab o Excel} = 1 – 0,8797 = 0,1203 (figura 32).

Figura 32. Probabilidades en una *F*-Snedecor



Nota

De forma análoga a como ocurría en el caso de las distribuciones binomial y normal, también existen tablas que permiten calcular, sin necesidad de utilizar software como Minitab o Excel, las probabilidades asociadas a una distribución *t*-Student o *F*-Snedecor (ver, p. ej., http://www.statsoft.com/textbook/distribution-tables).

Resumen

En este módulo se han presentado las técnicas básicas de la estadística descriptiva univariante: representación gráfica de datos discretos y continuos, organización de los datos mediante tablas de frecuencias y uso de estadísticos descriptivos para resumir datos. Conviene recordar que el tipo de gráfico, tabla o estadístico a usar dependerá siempre del tipo de variable considerada (categórica, cuantitativa discreta o cuantitativa continua), así como del tipo de información que se desee obtener.

Además, se ha explicado también el concepto de probabilidad de un suceso, que desempeña una función relevante en el análisis y predicción del comportamiento de las variables aleatorias asociadas a fenómenos cotidianos.

Finalmente, se han presentado algunos de los principales modelos matemáticos que se usan para describir, de forma teórica, el comportamiento de variables aleatorias: la distribución binomial, la normal, la *t*-Student y la *F*-Snedecor son algunos ejemplos de dichos modelos. El cálculo de probabilidades asociadas a variables que se comportan según alguno de estos modelos permite entender mejor su comportamiento y realizar estimaciones sobre la población de individuos de la que provienen los datos.

Ejercicios de autoevaluación

1) La tabla siguiente resume las respuestas ofrecidas por doscientos usuarios de un portal web a la pregunta "el nivel de usabilidad del portal es adecuado":

Respuesta	Frecuencia
Totalmente de acuerdo	50
De acuerdo	75
Ligeramente de acuerdo	25
Ligeramente en desacuerdo	15
En desacuerdo	15
Totalmente en desacuerdo	20

Se pide que hagáis lo siguiente:
a) Construir un diagrama de barras que permita visualizar las respuestas obtenidas.
b) Calcular la frecuencia relativa de aparición de cada respuesta y construir un diagrama circular para ilustrar dichos valores.

2) La tabla siguiente contiene cuarenta observaciones para el tiempo transcurrido (en horas) entre el envío de un mensaje a un foro en línea y su correspondiente respuesta.

4,0	3,5	3,1	6,0	5,6	3,1	2,9	3,8
4,3	3,8	4,5	3,5	4,5	6,1	2,8	5,0
5,4	3,8	6,8	4,9	3,6	3,6	3,8	3,7
4,1	2,0	3,7	5,7	7,8	4,6	4,8	2,8
5,0	5,2	4,0	5,4	4,6	3,8	4,0	2,9

A partir de estos datos, debéis hacer lo siguiente:
a) Construir un diagrama de tallos y hojas. Usad 1,0 como unidad de incremento.
b) Construir un histograma.
c) ¿Se observa en los datos algún patrón claro? ¿Cuál es la moda de la distribución de los datos?

3) La tabla siguiente muestra veinte observaciones de la variable aleatoria "número de correos electrónicos recibidos en un día".

3,9	3,4	5,1	2,7	4,4
7,0	5,6	2,6	4,8	5,6
7,0	4,8	5,0	6,8	4,8
3,7	5,8	3,6	4,0	5,6

Se pide que hagáis lo siguiente:
a) Hallar los estadísticos descriptivos de esta muestra. ¿Cuánto vale el rango intercuartílico? ¿Entre qué dos valores están comprendidos el 50% de los datos centrales de la muestra?
b) Construir un diagrama de cajas y bigotes (*boxplot*). ¿Hay algún valor anómalo (*outlier*) entre las observaciones?

4) Cuando se efectúa un control antidopaje a un atleta que no ha tomado sustancia alguna, la probabilidad de que el test dé un falso positivo es de 0,006. Si durante una competición se efectúa el test a un total de 1.000 atletas que están libres de sustancias, ¿cuál será el número esperado (promedio) de falsos positivos?, ¿cuál es la probabilidad de que el número de falsos positivos sea superior a quince?, ¿qué cabría pensar si aparecen más de quince positivos?

5) De acuerdo con el Instituto Nacional de Estadística, el 9,96% de los adultos residentes en España son extranjeros. Con el fin de realizar una encuesta, se pretende contactar con una muestra aleatoria de mil doscientos adultos residentes en España. ¿Cuál será el número espe-

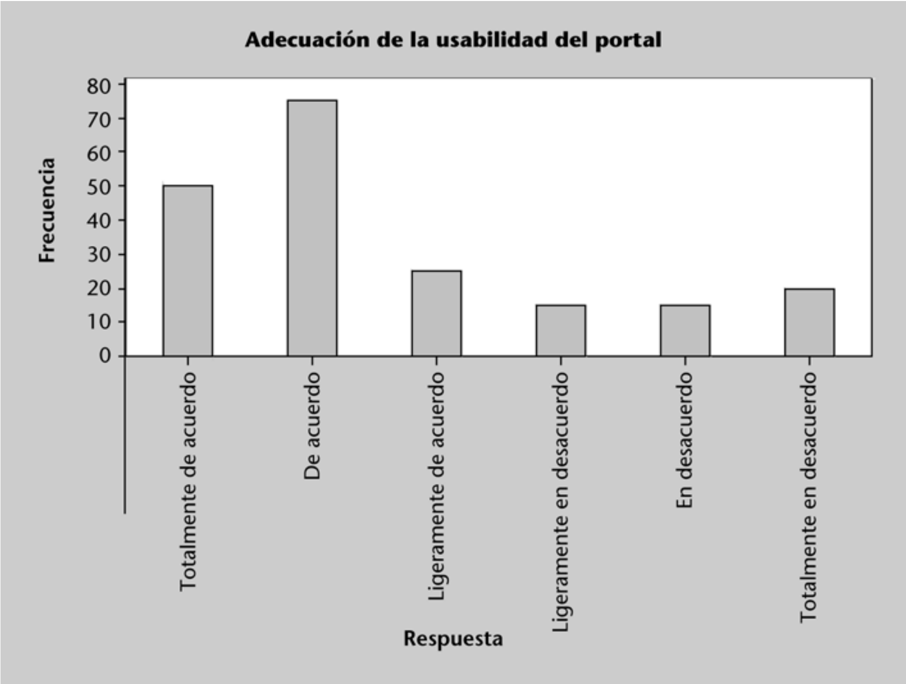
rado (promedio) de extranjeros que contendrá dicha muestra?, ¿cuál es la probabilidad de que la muestra contenga menos de cien extranjeros?

6) El tiempo de duración de un embarazo es una variable aleatoria que se distribuye de forma aproximadamente normal con una media de doscientos sesenta y seis días y una desviación estándar de dieciséis días. ¿Qué porcentaje de embarazos duran menos de doscientos cuarenta días (unos ocho meses)?, ¿qué porcentaje de embarazos duran entre doscientos cuarenta y doscientos setenta días (entre unos ocho y nueve meses)?, ¿a partir de cuántos días se sitúan el 20% de los embarazos más largos?

Solucionario

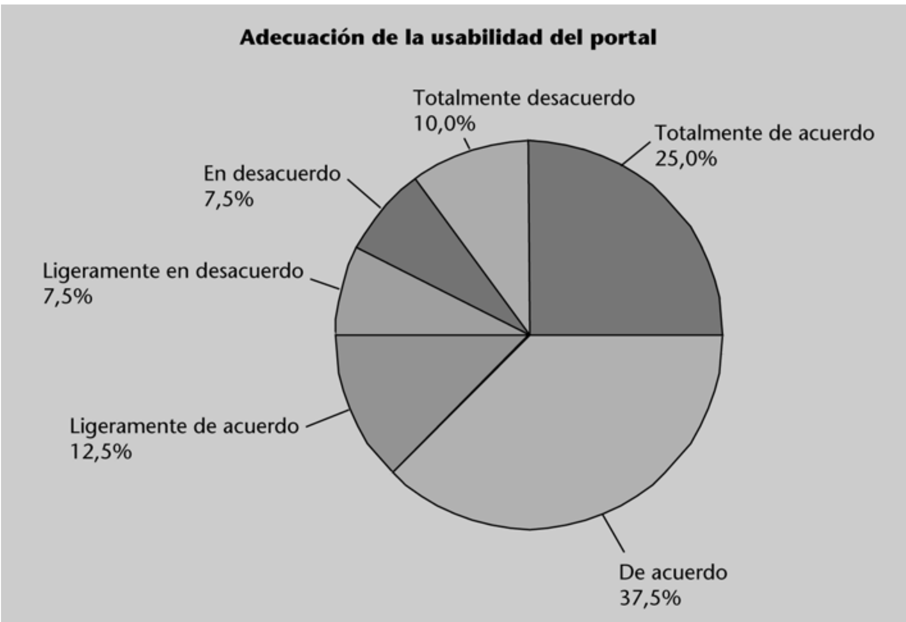
1)

a)



b)

Respuesta	Frecuencia	Frec. relativa
Totalmente de acuerdo	50	25,0%
De acuerdo	75	37,5%
Ligeramente de acuerdo	25	12,5%
Ligeramente en desacuerdo	15	7,5%
En desacuerdo	15	7,5%
Totalmente en desacuerdo	20	10,0%
Totales	**200**	**100%**



2)

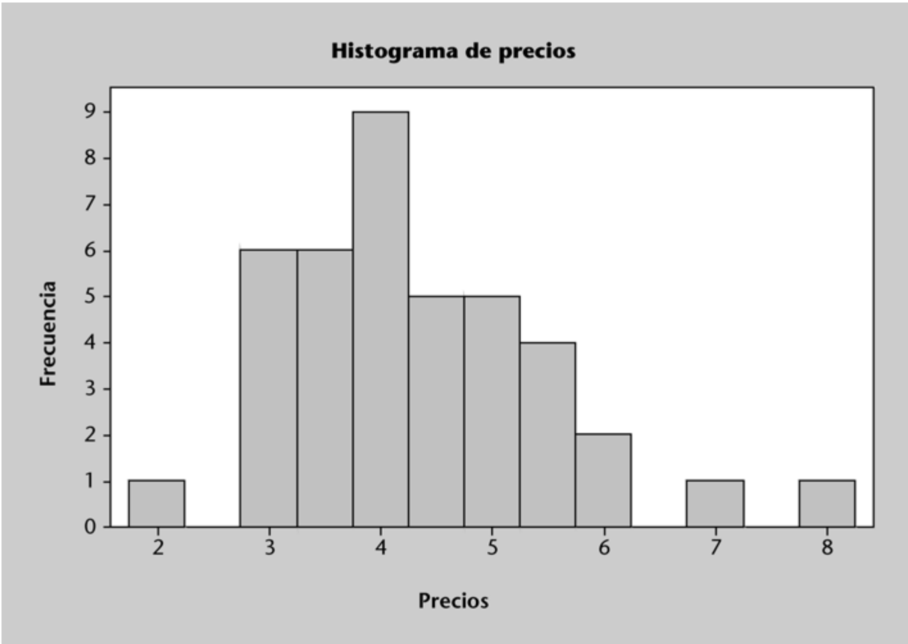
a)

```
Stem-and-Leaf Display: precios

Stem-and-leaf of precios  N  = 40
Leaf Unit = 0.10


 5    2  08899
 18   3  1155667788888
(11)  4  00013556689
 11   5  0024467
 4    6  018
 1    7  8
```

b)



c) Aunque no parece haber ningún patrón claro en los datos, sí se aprecia –tanto en el histograma como en el gráfico de tallos y hojas– una cierta forma de campana, con la parte central más elevada y unos extremos o colas más bajas. La moda de este conjunto de datos es 3,8 ya que, como se aprecia en el diagrama de tallos y hojas, es el valor que más aparece.

3)

a)

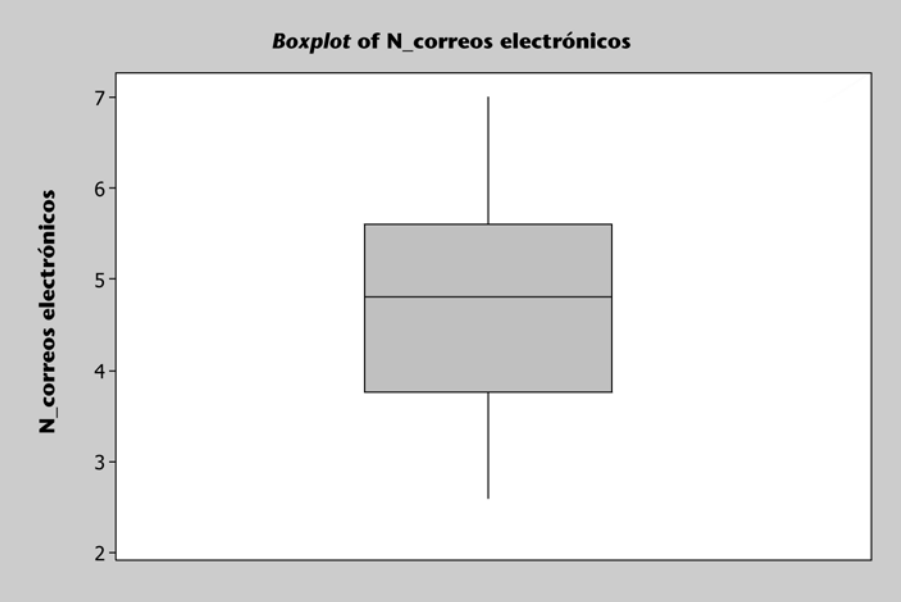
```
Descriptive Statistics: N_e-mails

Variable    N  N*   Mean  SE Mean  StDev  Minimum     Q1  Median     Q3
N_e-mails  20   0  4.810    0.291  1.302    2.600  3.750   4.800  5.600

Variable   Maximum
N_e-mails    7.000
```

El rango intercuartílico es Q3 – Q1 = 5,60 – 3,75 = 1,85. Entre Q1 = 3,75 y Q3 = 5,60 están comprendidos el 50% de los datos centrales.

b)



No se observa, en este caso, ningún valor anómalo (*outlier*), ya que el gráfico no muestra ningún símbolo "*".

4) En este caso, puesto que el resultado de cada test puede ser "positivo" (con probabilidad 0,006) o "no positivo" (con probabilidad 1 – 0,006 = 0,994), la variable aleatoria X = "número de falsos positivos en 1.000 pruebas a atletas limpios" sigue una distribución binomial de parámetros n = 1.000 y p = 0,006. En el caso de la binomial, la media o valor esperado es $\mu = n \cdot p = 6$, es decir, cabe esperar que al aplicar el test a 1.000 atletas "limpios" haya seis falsos positivos.

Por otra parte, $P(X > 15) = 1 - P(X <= 15)$ = {Minitab o Excel} = 1 – 0,9995 = 0,0005. Por tanto, si aparecen más de quince positivos cabría pensar que muy probablemente no todos ellos sean falsos.

5) En este caso, la variable aleatoria X = "número de extranjeros en la muestra" sigue una distribución binomial de parámetros n = 1.200 y p = 0,0996. Por tanto, el valor esperado de extranjeros en la muestra es $\mu = n \cdot p$ = 119,52, es decir el promedio de extranjeros para las muestras de esas características es de, aproximadamente, 120.

Por otro lado, $P(X < 100) = P(X <= 99)$ = {Minitab o Excel} = 0,0245, es decir, es muy poco probable que una muestra contenga menos de 100 extranjeros si ésta es realmente aleatoria.

6) Se considera la variable aleatoria X = "días que dura un embarazo". Cabe tener en cuenta que $X \sim N$ (266,16).

$P(X < 240)$ = {Minitab o Excel} = 0,0521, es decir, el 5,2% de los embarazos duran menos de ocho meses.

$P(240 < X < 270) = P(X < 270) - P(X < 240)$ = {Minitab o Excel} = 0,5987 – 0,0521 = 0,5466, es decir, el 55% de los embarazos duran entre ocho y nueve meses.

Finalmente, se pide el valor c tal que $P(X > c)$ = 0,20, es decir: $P(X < c) = 1 - P(X > c)$ = 0,80 → c = {Minitab o Excel} = 279,47, es decir, el 20% de los embarazos supera los doscientos setenta y nueve días.

Inferencia de información para una población

Distribuciones muestrales y teorema central del límite. Intervalos de confianza. Contrastes de hipótesis para una población

Blanca de la Fuente

PID_00161059

Universitat Oberta de Catalunya
www.uoc.edu

Índice

Introducción

El objetivo de la inferencia estadística es obtener información acerca de una población, partiendo de la información que contiene la muestra. La selección de la muestra debe garantizar su representatividad, lo que se consigue eligiéndola al azar mediante diferentes procedimientos de muestreo que se estudian en el módulo 5.

Una vez seleccionada una muestra, se dispone de un conjunto de valores, en cuyo caso los *métodos descriptivos* estudiados en el módulo 1 facilitan el análisis de estos valores muestrales. El problema que ahora se aborda es la extensión de estos resultados al conjunto de la población o, en otras palabras, dar respuesta al siguiente interrogante: Dada cierta información muestral ¿qué podemos afirmar de la población?

La solución de este problema será el objetivo de la *inferencia estadística*.

Hasta ahora se había supuesto que los valores de los parámetros de las distribuciones de probabilidad eran conocidos. Pero esto casi nunca ocurre, de manera que tenemos que usar los datos muestrales para estimarlos. Los **estimadores** proveen valores a esos parámetros.

Cuando las inferencias que se realizan se refieren a características poblacionales concretas, es necesaria una etapa de diseño de estimadores. En este módulo se presentan los conceptos básicos para la estimación de la proporción, de la media y de la varianza de la población respectivamente.

Un enfoque alternativo es indicar un rango de valores, entre los cuales tiene que estar el parámetro con una determinada precisión: esta es la idea de un **intervalo de confianza**.

A continuación se plantea en este módulo el problema del **contraste de hipótesis**, desarrollando métodos que permiten contrastar la validez de una conjetura o de una afirmación utilizando datos muestrales. El proceso comienza cuando un investigador formula una hipótesis sobre la naturaleza de una población. La formulación de esta hipótesis implica claramente la elección entre dos opciones; a continuación, el investigador selecciona una opción basándose en los resultados de un estadístico calculado a partir de una muestra aleatoria de datos.

Objetivos

Los objetivos académicos del presente módulo se describen a continuación:

1. Explorar las distribuciones de la media, de la proporción y de la varianza muestral.

2. Aplicar el Teorema central del límite.

3. Crear intervalos de confianza.

4. Usar la distribución *t* en una prueba de hipótesis.

5. Utilizar la distribución chi-cuadrado (χ^2) en una prueba de hipótesis.

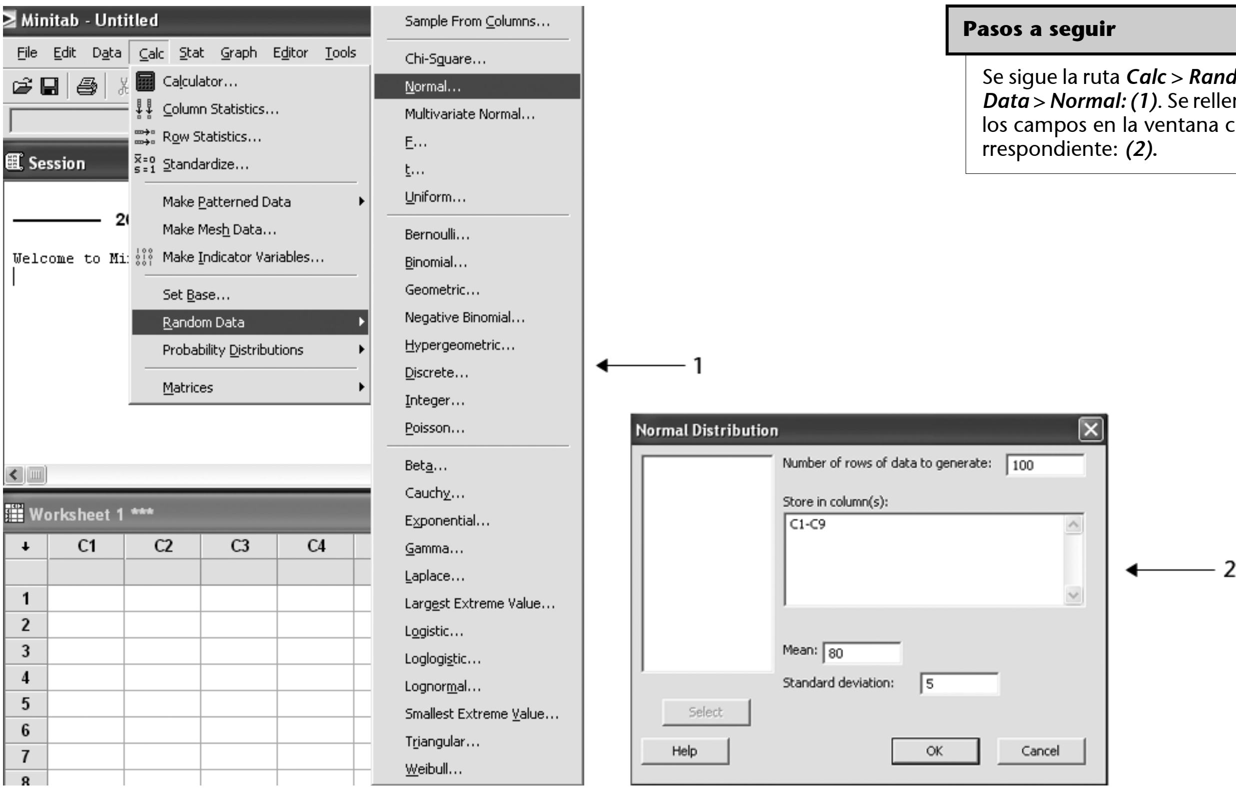
1. Distribuciones muestrales y Teorema central del límite

Una muestra aleatoria permite hacer inferencia sobre ciertas características de la distribución de la población. Esta inferencia estará basada en algún **estadístico**, es decir, alguna función particular de la información muestral. La **distribución muestral** de este estadístico es la distribución de probabilidades de los valores que puede tomar el estadístico a lo largo de todas las posibles muestras con el mismo número de observaciones, que pueden ser extraídas de la población.

Por ejemplo, en la distribución normal, los dos parámetros son la media de la población μ y la desviación estándar poblacional σ. Se puede estimar el valor μ calculando el promedio muestral o media muestral, $\bar{x}$, y el valor de σ mediante el cálculo de la desviación típica muestral, s. En este caso la media muestral, $\bar{x}$ y la desviación típica muestral, s, son los estadísticos. En el caso de la distribución binomial, los parámetros son n y p. Para estimar el parámetro proporción poblacional, p, se utiliza el estadístico proporción muestral, $\hat{p}$.

El estudio de las distribuciones muestrales se puede ilustrar mediante la creación con Minitab de 100 muestras de datos aleatorios normales con media 80 y desviación típica 5, con 9 observaciones de cada muestra (figura 1). A partir de datos aleatorios se crea una columna de datos que contenga el promedio de cada muestra o media muestral.

Figura 1. Pasos a seguir para estudiar una distribución muestral

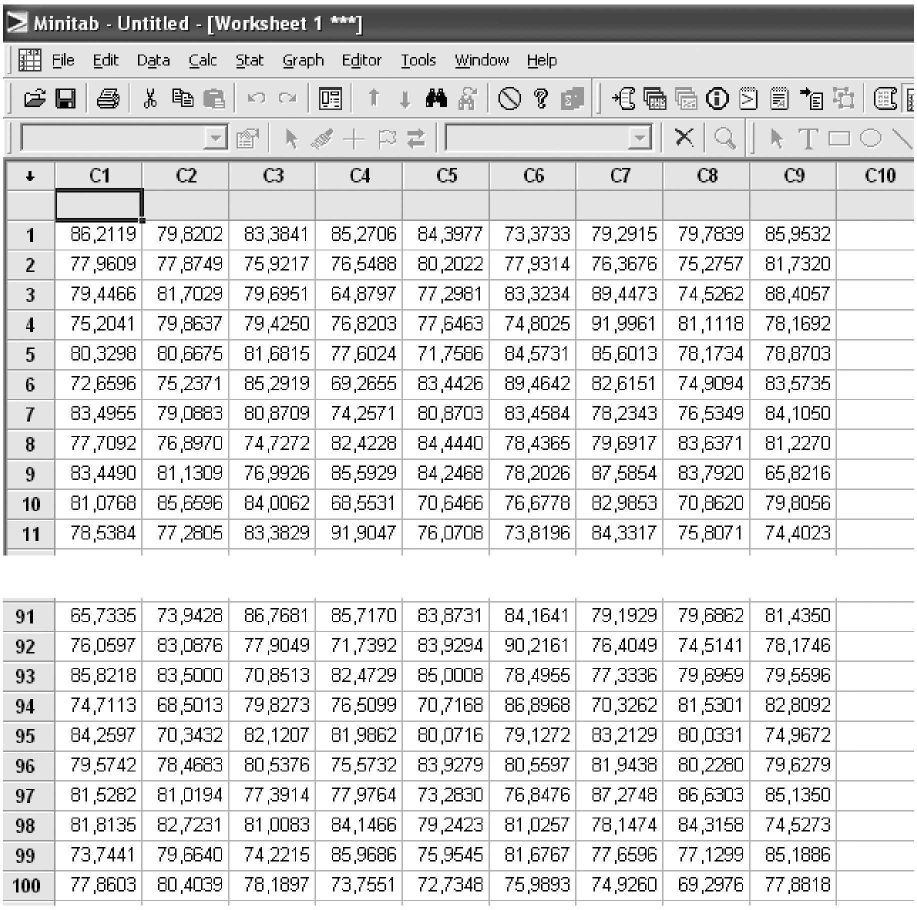


Pasos a seguir

Se sigue la ruta ***Calc*** > ***Random Data*** > ***Normal: (1)***. Se rellenan los campos en la ventana correspondiente: ***(2).***

Se ha generado así una matriz de nueve columnas y cien filas (figura 2). Cada componente de esta matriz es una observación aleatoria proveniente de una distribución normal de media 80 y desviación estándar 5.

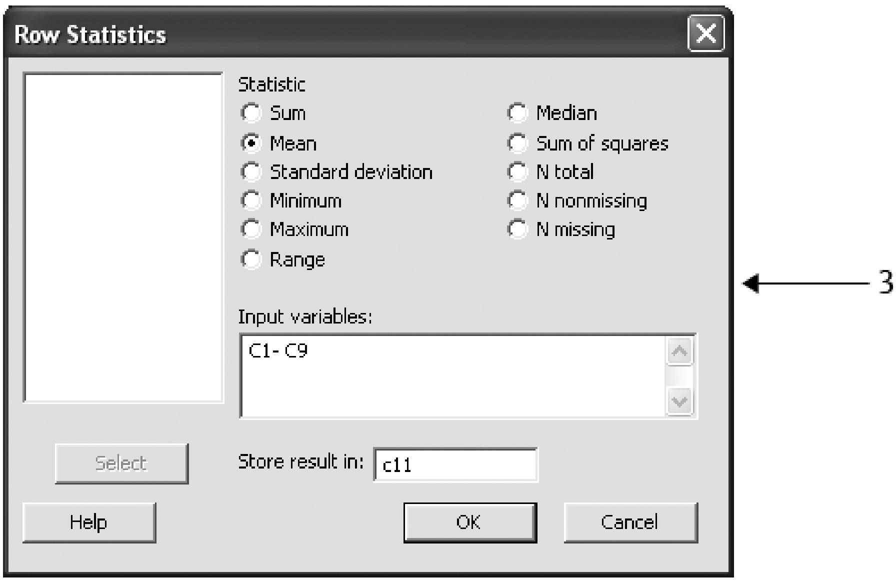
Figura 2. Resultado de una matriz



↓	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10
1	86,2119	79,8202	83,3841	85,2706	84,3977	73,3733	79,2915	79,7839	85,9532	
2	77,9609	77,8749	75,9217	76,5488	80,2022	77,9314	76,3676	75,2757	81,7320	
3	79,4466	81,7029	79,6951	64,8797	77,2981	83,3234	89,4473	74,5262	88,4057	
4	75,2041	79,8637	79,4250	76,8203	77,6463	74,8025	91,9961	81,1118	78,1692	
5	80,3298	80,6675	81,6815	77,6024	71,7586	84,5731	85,6013	78,1734	78,8703	
6	72,6596	75,2371	85,2919	69,2655	83,4426	89,4642	82,6151	74,9094	83,5735	
7	83,4955	79,0883	80,8709	74,2571	80,8703	83,4584	78,2343	76,5349	84,1050	
8	77,7092	76,8970	74,7272	82,4228	84,4440	78,4365	79,6917	83,6371	81,2270	
9	83,4490	81,1309	76,9926	85,5929	84,2468	78,2026	87,5854	83,7920	65,8216	
10	81,0768	85,6596	84,0062	68,5531	70,6466	76,6778	82,9853	70,8620	79,8056	
11	78,5384	77,2805	83,3829	91,9047	76,0708	73,8196	84,3317	75,8071	74,4023	
91	65,7335	73,9428	86,7681	85,7170	83,8731	84,1641	79,1929	79,6862	81,4350	
92	76,0597	83,0876	77,9049	71,7392	83,9294	90,2161	76,4049	74,5141	78,1746	
93	85,8218	83,5000	70,8513	82,4729	85,0008	78,4955	77,3336	79,6959	79,5596	
94	74,7113	68,5013	79,8273	76,5099	70,7168	86,8968	70,3262	81,5301	82,8092	
95	84,2597	70,3432	82,1207	81,9862	80,0716	79,1272	83,2129	80,0331	74,9672	
96	79,5742	78,4683	80,5376	75,5732	83,9279	80,5597	81,9438	80,2280	79,6279	
97	81,5282	81,0194	77,3914	77,9764	73,2830	76,8476	87,2748	86,6303	85,1350	
98	81,8135	82,7231	81,0083	84,1466	79,2423	81,0257	78,1474	84,3158	74,5273	
99	73,7441	79,6640	74,2215	85,9686	75,9545	81,6767	77,6596	77,1299	85,1886	
100	77,8603	80,4039	78,1897	73,7551	72,7348	75,9893	74,9260	69,2976	77,8818	

Se considera que cada una de las filas obtenidas es una muestra, y se calcula la media asociada a cada una de estas cien muestras (figura 3):

Figura 3. Pasos a seguir para calcular las medias



Pasos a seguir

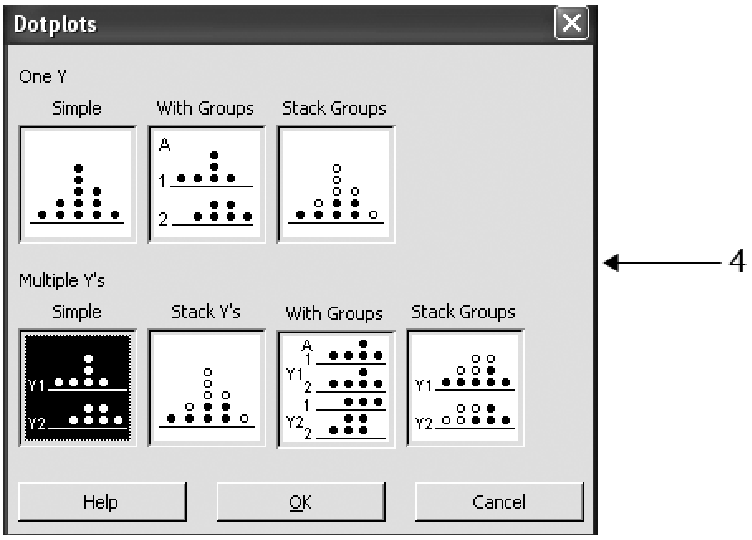
Una vez generados los datos se sigue la ruta ***Calc*** > ***Row Statistics*** y se rellenan los campos en la ventana correspondiente: ***(3)***.

En la columna C11 de la figura 4 hay cien nuevos valores (las medias). En la figura 5 se muestran los *dotplot* asociados a las columnas C1 (que representan cien valores aleatorios obtenidos de una normal 80-5) y C11:

Figura 4. Resultado del análisis

↓	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12
											x-barra	
1	86,2119	79,8202	83,3841	85,2706	84,3977	73,3733	79,2915	79,7839	85,9532		81,9429	
2	77,9609	77,8749	75,9217	76,5488	80,2022	77,9314	76,3676	75,2757	81,7320		77,7572	
3	79,4466	81,7029	79,6951	64,8797	77,2981	83,3234	89,4473	74,5262	88,4057		79,8583	
4	75,2041	79,8637	79,4250	76,8203	77,6463	74,8025	91,9961	81,1118	78,1692		79,4488	
5	80,3298	80,6675	81,6815	77,6024	71,7586	84,5731	85,6013	78,1734	78,8703		79,9175	
6	72,6596	75,2371	85,2919	69,2655	83,4426	89,4642	82,6151	74,9094	83,5735		79,6065	
7	83,4955	79,0883	80,8709	74,2571	80,8703	83,4584	78,2343	76,5349	84,1050		80,1016	
8	77,7092	76,8970	74,7272	82,4228	84,4440	78,4365	79,6917	83,6371	81,2270		79,9103	
9	83,4490	81,1309	76,9926	85,5929	84,2468	78,2026	87,5854	83,7920	65,8216		80,7571	
10	81,0768	85,6596	84,0062	68,5531	70,6466	76,6778	82,9853	70,8620	79,8056		77,8081	
11	78,5384	77,2805	83,3829	91,9047	76,0708	73,8196	84,3317	75,8071	74,4023		79,5042	
12	70,0951	78,7096	77,3802	82,9569	72,4905	76,7535	88,9660	85,7643	75,4986		78,7350	

Figura 5. Pasos a seguir para crear el gráfico de puntos de los *dotplot*



Pasos a seguir

Se sigue la ruta ***Graph*** > ***Dotplot*** y se rellenan los campos en la ventana correspondiente: ***(4)***.

La salida de Minitab de la figura 6 muestra que la distribución de la variable aleatoria inicial *X* (columna C1) era normal y, según el gráfico de puntos, parece que también la distribución de la v.a. *X*-barra ($\bar{x}$) es normal, de media muy similar y desviación estándar menor (los puntos de la $\bar{x}$ están menos "dispersos" que los de la *x*).

También podemos hacer un histograma de frecuencias de la distribución de las medias muestrales ($\bar{x}$), como se aprecia en la figura 7.

Figura 6. Gráfico de puntos de valores de los *dotplot*

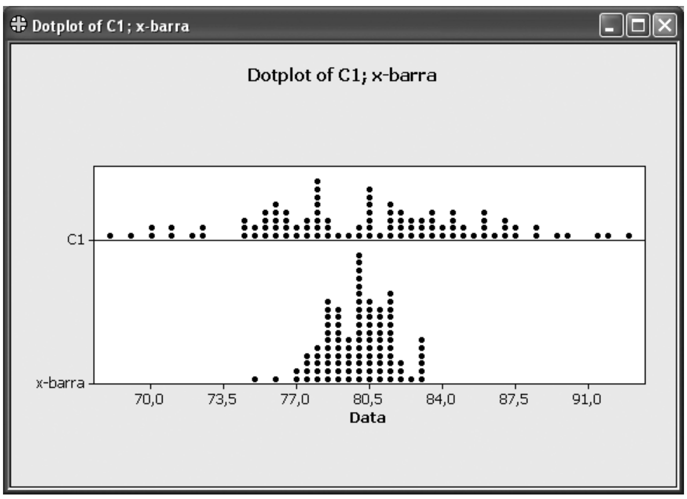
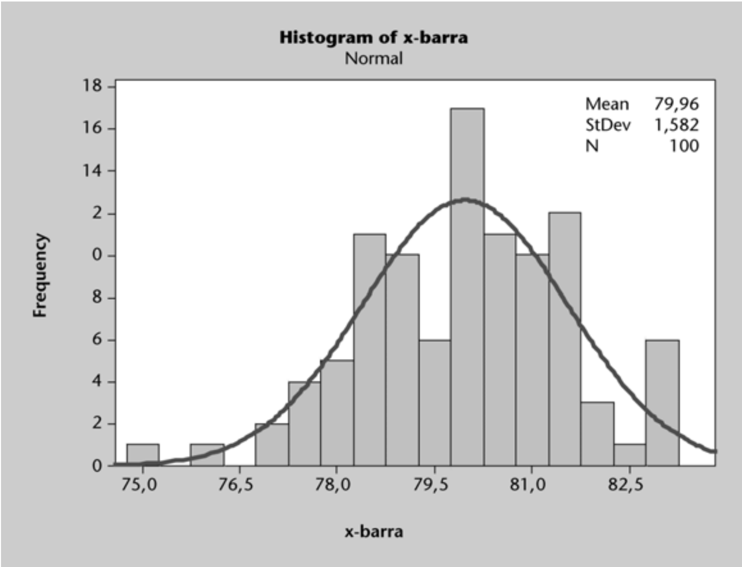


Figura 7. Histograma de frecuencias absolutas de valores de $\bar{x}$ a partir de nueve muestras aleatorias simples, cada una de tamaño cien



Finalmente, en la figura 8 se obtienen los estadísticos que describen la distribución de las medias muestrales.

Figura 8. Resultado del análisis de *X*-barra

Descriptive Statistics: x-barra

```
Variable    N  N*    Mean  SE Mean  StDev  Minimum      Q1
Median      Q3
x-barra   100   0  79,962    0,158  1,582   75,192  78,814
80,146  81,000

Variable  Maximum
x-barra    83,154
```

Pasos a seguir

Se sigue la ruta ***Stat*** > ***Basic Statistics*** > ***Display Descriptive Statistics*** y se selecciona la variable ***C11*** (***x-barra***) en la ventana correspondiente.

La media de los cien valores contenidos de la columna C11 (y que es una aproximación a la media de la v.a. *X*-barra) es de 79,962, valor muy similar a la media de *X* (que era de 80). Esto es coherente con lo que la teoría nos indica:

- La media muestral coincide con la media de la población, $\mu_{\bar{X}} = \mu$.

La desviación estándar de los cien valores de la columna C11 (que será una aproximación a la desviación estándar de *X*-barra) es de 1,582. Si tomamos la desviación estándar de *X* (que era de 5) y la dividimos por 3 (raíz de 9, el tamaño de la muestra), obtenemos el valor 1,667.

- Ambos valores son muy parecidos, tal y como la teoría predice:

$$\sigma_{\bar{X}} = \sigma / \sqrt{n}$$

Es interesante señalar que si no se hubiera tomado inicialmente una variable normalmente distribuida, las conclusiones obtenidas serían semejantes siempre que el tamaño muestral *n* fuera lo suficientemente grande tal y como predice el **Teorema central del límite.**

Teorema central del límite

El análisis anterior se aplica sólo a la distribución normal. ¿Qué ocurre si nuestros datos provienen de otra distribución de probabilidad? ¿Podemos decir algo acerca de la distribución muestral de la media en ese caso? Para ello se utiliza el ***Teorema central del límite,*** el cual expresa que si tenemos una muestra tomada de una distribución de probabilidad con media μ y desviación típica de σ, la distribución muestral de $\bar{x}$ es aproximadamente normal con media μ y desviación típica de, $\sigma/\sqrt{n}$ que es el error estándar. Lo notable acerca del teorema central del límite es que la distribución de la media muestral de $\bar{x}$ es más o menos normal, sea cual sea la distribución original de probabilidad. A medida que aumenta el tamaño de la muestra, la aproximación a la distribución normal se acerca cada vez más.

Nota

Consideraremos que n es lo bastante grande cuando, como mínimo, $n > 30$.

Una consecuencia de este teorema es:

Dada cualquier variable aleatoria con esperanza μ y para n suficientemente grande, la distribución de la variable:

$$\frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$$

es una normal estándar $N(0,1)$.

Cálculo del error estándar

Recordemos que si la variable tiene una desviación típica conocida σ, el error estándar se puede calcular como $\sigma/\sqrt{n}$. Cuando σ es desconocida, calculamos el error estándar como $s/\sqrt{n}$, siendo s la desviación típica de la muestra.

Un caso particular es la **aproximación de la binomial a la normal**:

Sea X una variable aleatoria con distribución $B(n, p)$ binomial con n suficientemente grande. Entonces, X es aproximadamente normal con esperanza np y varianza $np(1-p)$.

En este caso, n grande significa que np y $np(1-p)$ son los dos mayores que 5 o bien que $n > 30$.

Por tanto, cuando el tamaño de la muestra, n, es grande, la distribución de la **proporción** es aproximadamente una distribución normal de esperanza p y desviación típica $\sqrt{p(1-p)/n}$. En este caso $\sqrt{p(1-p)/n}$, corresponde al error estándar $N\left(p,\sqrt{\frac{p(1-p)}{n}}\right)$.

Recordatorio

Si X sigue una distribución **binomial** de parámetros n y p, entonces:

$$P(X=k)=\binom{n}{k}p^k(1-p)^{n-k}$$

para los $k \in \{0, \ldots, n\}$

Ejemplo: se hace una encuesta sobre un determinado tema que tiene dos opciones, A y B. La probabilidad de que un individuo concreto opine A es p y n es el número de encuestas hechas. Hemos preguntado a cuatrocientos habi-

tantes y encontramos que el 30% opina A, es decir, que podemos establecer que $p = 0,3$. Entonces, la distribución de la proporción de habitantes que opina A sigue una distribución normal, cuya media es 0,3, que coincide con la proporción del 30% de los habitantes de la población que opinan A, y la desviación estándar es 0,0229, que corresponde a la desviación típica de la población dividida por la raíz cuadrada del tamaño de la muestra.

$$N\left(0,3, \sqrt{\frac{0,3(1-0,3)}{400}}\right) = N(0,3; 0,0229)$$

2. Distribución de la media muestral

Se deben considerar dos casos para la distribución de la media muestral.

Caso de desviación típica poblacional conocida

Si la variable que estudiamos sigue una distribución normal con media μ y desviación típica σ conocidas, entonces la media muestral es también normal con la misma media μ y desviación típica $\sigma/\sqrt{n}$, donde n es el tamaño de la muestra.

Nota

Si σ es la desviación típica de la población y n el tamaño de la muestra, se define el **error estándar de la media muestral** como:

$$\sigma/\sqrt{n}$$

Siempre que la distribución de las medias muestrales sea una distribución normal, se puede calcular una **variable aleatoria normal estandarizada, *Z***, que tiene una media 0 y una varianza 1:

Observad

El error estándar es cada vez menor cuanto mayor es el tamaño de la muestra.

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

Si la distribución de la población no es normal pero el tamaño muestral n es suficientemente grande, entonces se usará el teorema central del límite y la variable media muestral se aproxima a una normal estándar a medida que el tamaño de la muestra aumenta. En general, dicha aproximación se considera válida para tamaños muestrales superiores a treinta.

En el apartado anterior se vio que la variable aleatoria binomial sigue una distribución normal aproximada cuando aumenta el tamaño de la muestra.

Ejemplo: en la asignatura de *Archivística* de una licenciatura de Documentación se sabe que las calificaciones siguen una distribución normal de media 7,4 y desviación estándar 0,78. Se desea conocer el porcentaje de estudiantes con nota superior a 6,5 e inferior a 8,5. ¿Con qué nota se va a calificar como "excelente" (A), si esta es la calificación del 5% de estudiantes con mejor nota?

Solución:

La variable sigue una distribución $N(7{,}4;\ 0{,}78)$. Primero se calcula el estadístico Z normal estandarizado:

$$P(6{,}5 \le X \le 8{,}5) = P\left(\frac{6{,}5-7{,}4}{0{,}78} \le \frac{X-7{,}4}{0{,}78} \le \frac{8{,}5-7{,}4}{0{,}78}\right) =$$
$$= P(-1{,}15 \le Z \le 1{,}41) =$$
$$= P(Z \le 1{,}41) - P(Z \le -1{,}15) = 0{,}9207 - 0{,}1251 = 0{,}7956$$

Los valores de probabilidad se buscan en la tabla $N(0,1)$ o calculándose con cualquier programa estadístico como se muestra en el ejemplo desarrollado en el módulo 1.

A la vista del resultado, se puede decir que el porcentaje de estudiantes con nota superior a 6,5 e inferior a 8,5 es de 79,56%.

Para calcular la nota a partir de la cual se califica como excelente, se calcula el estadístico Z normal estandarizado:

$$P(X \geq A) = P\left(\frac{X-7,4}{0,78} \geq \frac{A-7,4}{0,78}\right) = P(Z \geq z_A) = 0,05$$

En las tablas de la $N(0,1)$ o mediante cualquier programa estadístico se busca un valor z que deje a la derecha un área de 0,05, aproximadamente el valor es: $z_A = 1,645$, de manera que:

$$\frac{A-7,4}{0,78} = 1,645 \qquad \Rightarrow \qquad A = 7,4 + 1,645 \cdot 0,78 = 8,683$$

A partir de una nota de 8,6 se califica como "excelente"(A).

Caso de desviación típica poblacional desconocida

Cuando la desviación poblacional es desconocida y el tamaño de la muestra es pequeño, deberemos hacer una estimación de la desviación típica con la llamada *desviación típica muestral*. Para ello es necesario presentar una nueva distribución de probabilidad. Esta nueva distribución se conoce con el nombre de ***t* de Student** cuyas características se explicaron en el módulo 1.

Para determinar la distribución de la media muestral cuando la desviación poblacional es desconocida, se debe calcular la desviación típica muestral:

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

Si la variable estudiada sigue una distribución normal con media μ y desviación típica desconocida, entonces el estadístico media muestral sigue una distribución t_{n-1}, es decir, una ***t* de Student con *n*–1 grados de libertad.**

Nota

En este caso se define el **error estándar de la media muestral** como:

$$s/\sqrt{n}$$

$$t_{n-1} = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

Los grados de libertad asociados con el valor de t son n–1(tamaño de la muestra menos uno).

Ejemplo: el tiempo que han tardado en infectarse de virus cada uno de los ordenadores de una editorial ha sido: 2,5; 7,4; 8,0; 4,5; 7,4 y 9,2 segundos.

Suponemos que el tiempo que tarda un ordenador de esa editorial en infectarse sigue la distribución normal de media 6,5 y se desconoce la varianza poblacional. Se desea calcular la probabilidad de que un ordenador tarde entre 5 y 10 segundos en infectarse.

Solución:

Como se desconoce la varianza de la población, la media muestral seguirá una distribución ***t* de Student con 5 grados de libertad**.

Para calcular el valor del estadístico *t*, se debe calcular la desviación típica muestral. El valor obtenido es S = 2,5:

$$t_{n-1} = \frac{\overline{x} - \mu}{\frac{s}{\sqrt{n}}}$$

La probabilidad solicitada será:

$$p(5 \leq T \leq 10) = p\left(\frac{5-6,5}{2,5/\sqrt{6}} \leq t_5 \leq \frac{10-6,5}{2,5/\sqrt{6}}\right) = p(-1,47 \leq t_5 \leq 3,43) =$$
$$p(t_5 \leq 3,43) - p(t_5 \leq -1,47) = 0,99 - 0,1 = 0,89$$

Para calcular la probabilidad se utiliza la tabla *t* o un programa estadístico (figura 9).

Figura 9. Resultado de Minitab

Cumulative Distribution Function

```
Student's t distribution with 5 DF

   x  P( X <= x )
3,43     0,990682
```

Cumulative Distribution Function

```
Student's t distribution with 5 DF

    x  P( X <= x )
-1,47     0,100758
```

Pasos a seguir

Para calcular las probabilidades de una distribución *t* de Student se sigue la ruta ***Calc*** > ***Probability Distributions*** > ***t*** y se completan los parámetros en la ventana correspondiente. El resultado se muestra en la figura 9.

3. Distribución de la proporción muestral

En el apartado 5 del módulo 1 se dijo que la distribución binomial era la suma de n variables aleatorias independientes, cada una de las cuales tiene una probabilidad de éxito p. Para caracterizar la distribución se necesita conocer el valor de p, que es la proporción de miembros de la población que tienen una característica de interés. La **proporción muestral de éxitos** en una muestra aleatoria extraída de una población en la que la proporción de éxitos p será:

Distribución de la proporción muestral

Es una aplicación del **Teorema central del límite**.

$$\hat{p} = \frac{X}{n}$$

Por lo tanto $\hat{p}$ es la media de un conjunto de variables aleatorias independientes. Además puede utilizarse el teorema central del límite para sostener que la distribución de probabilidad de $\hat{p}$ puede considerarse una distribución normal si el tamaño de la muestra es grande.

$$N\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$$

Igual que en el caso de la media muestral, siempre que la distribución de la proporción muestral sea una distribución normal, se puede calcular una **variable aleatoria normal estandarizada**, **Z**, que tiene una media cero y una varianza uno.

Nota

La distribución de $\hat{p}$ tiene una media igual a la proporción poblacional p.
La desviación estándar de $\hat{p}$ es el **error estándar de la media muestral** como:

$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$$

$$z = \frac{\hat{p} - p}{\sigma_{\hat{p}}}$$

La proporción muestral tiene muchas aplicaciones, entre las cuales se encuentra el estudio de los resultados de encuestas, la estimación de la cuota porcentual del mercado, el porcentaje de inversiones empresariales que tiene éxito y los resultados electorales entre otros.

Observad

El error estándar es cada vez menor cuanto mayor es el tamaño de la muestra.

Ejemplo: el 22% de los discos se venden por la Red en formato MP3 y el resto se vende en tiendas en formato CD. Se consideran las ventas de los próximos 5.000 discos. Se desea saber ¿qué distribución sigue la proporción muestral de discos vendidos por la Red? ¿Cuál es el número esperado de discos que se venderán por la Red? ¿Cuál es la probabilidad de que se vendan por la Red más de 1.500 discos?

Solución:

En este ejercicio se tiene que $p = 0{,}22$ y $n = 5.000$.

Para determinar la distribución de la proporción muestral, dado que el tamaño de la muestra es grande $n = 5.000$, se aplica el teorema central del límite. La distribución será aproximadamente **normal**, el valor de la media es el de la proporción poblacional (0,22).

Se calculará el error estándar $s_{\hat{p}} = \sqrt{\dfrac{0,22(1-0,22)}{5.000}} = 0,00586$

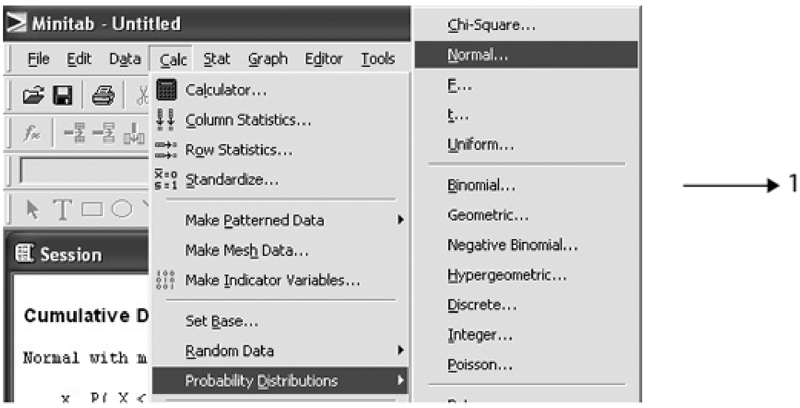
El valor esperado de discos vendidos por la Red será del 22% de los 5.000 que se venden en total, es decir, 1.100 discos en formato MP3.

La probabilidad de que se vendan menos de 1.500 discos por la Red será igual a la probabilidad de que la proporción muestral sea superior o igual al 30%. Para obtener esta probabilidad, primero se calculará el estadístico Z normal estandarizado:

$$P(p > 30\%) = P\left(Z > \frac{0,30 - 0,22}{0,00586}\right) = P(Z > 13,41) = 0$$

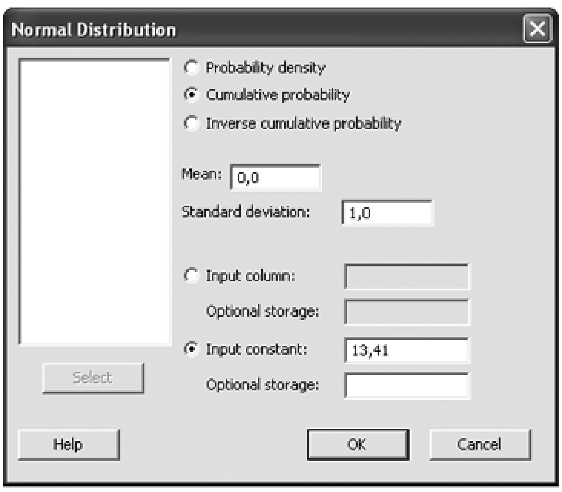
La probabilidad de Z se obtiene en la tabla $N(0,1)$. En la práctica, los cálculos probabilísticos anteriores se suelen automatizar con la ayuda de algún software estadístico o de análisis de datos. La figura 10 muestra cómo se pueden calcular probabilidades de una normal con ayuda de Minitab.

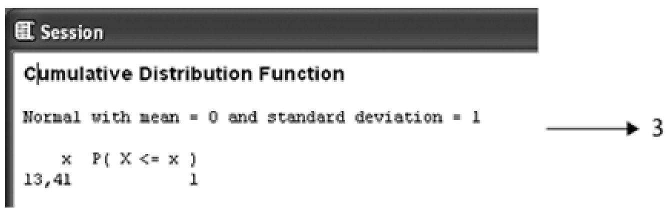
Figura 10. Cálculo de probabilidades con Minitab



Pasos a seguir

Se sigue la ruta ***Calc*** > ***Probability Distributions*** > ***normal*** ***(1)*** y se completan los parámetros en la ventana correspondiente ***(2)***. El resultado se muestra en ***(3)***. El programa calcula $P(Z <= 13,41)$.





El valor obtenido con Minitab es $P(Z <= 13{,}41)$. Por lo tanto, para obtener la probabilidad deseada calcularemos la probabilidad complementaria $P(Z > 13{,}41) = 1 - P(Z <= 13{,}41) = 1 - 1 = 0$.

4. Distribución de la varianza muestral

Una vez analizadas las distribuciones de las medias muestrales y las proporciones muestrales, se examinarán las distribuciones de las varianzas muestrales. A medida que las empresas y la industria ponen más énfasis en la producción de productos que satisfagan los criterios de calidad, es mayor la necesidad de calcular y reducir la varianza poblacional. Cuando la varianza es alta en un proceso, algunas características de los productos pueden tener una gama más alta de valores, como consecuencia de la cual hay más productos que no tienen un nivel de calidad aceptable. Se pueden obtener productos de calidad si el proceso de producción tiene una varianza baja, de manera que es menor el número de unidades que tienen un nivel de calidad inferior al deseado. Comprendiendo la distribución de las varianzas muestrales podemos hacer inferencias sobre la varianza poblacional.

Si se estudia una muestra aleatoria de tamaño n y varianza muestral s^2 obtenida de una población normal de media μ y varianza σ^2 desconocidas, entonces la varianza muestral se distribuye como una χ^2_{n-1} con n–1 grados de libertad:

$$\chi^2_{n-1} = \frac{(n-1)\,s_x^2}{\sigma_x^2}$$

Por lo tanto, se pueden hacer inferencias sobre la varianza poblacional σ^2 utilizando s^2 y la distribución chi-cuadrado. Este proceso se muestra en el siguiente ejemplo.

Ejemplo: en una gran ciudad se ha observado que durante el verano las facturas del consumo de electricidad siguen una distribución normal que tiene una desviación típica del 100 euros. Se ha tomado una muestra aleatoria de 25 facturas. Se desea calcular la probabilidad de que la desviación típica muestral sea inferior a 75 euros.

Solución:

En este ejercicio se tiene que $n = 25$ y $\sigma^2 = (100)^2$. Utilizando la distribución chi-cuadrado se puede establecer que:

$$P\left(s^2 < 75^2\right) = P\left(\frac{(25-1)75^2}{(100)^2} < \chi^2_{24\,g.l.}\right) = P\left(13{,}5 < \chi^2_{24\,g.l.}\right)$$

Los valores de la distribución chi-cuadrado pueden obtenerse en la tabla de dicha distribución con 24 grados de libertad:

$$\chi^2_{24\,g.l.} = 12{,}401;\ \chi^2_{24\,g.l.} = 13{,}848$$

El valor de probabilidad estará entre 0,025 y 0,05 (0,0428) exactamente.

5. Intervalos de confianza para una población

En los apartados anteriores hemos considerado la estimación puntual de un parámetro desconocido de la población, es decir, el cálculo de un único número que sea una buena aproximación. En la mayoría de los problemas prácticos, un estimador puntual por sí solo es inadecuado. Por ejemplo, supongamos que un control hecho sobre una muestra aleatoria de manuales procedentes de un gran envío de una editorial nos lleva a estimar que el 10% de todos los manuales son defectuosos. Un gerente que se enfrenta a este dato posiblemente se hará preguntas del tipo: ¿puede estar totalmente seguro de que el verdadero valor del porcentaje de manuales defectuosos está entre el 5% y el 15%? O ¿es muy posible que entre el 9% y el 11% de los manuales sean defectuosos? Esta clase de preguntas requieren información que va más allá de la contenida en una simple estimación puntual; son preguntas que buscan la fiabilidad de dicho estimador. En otras palabras, se trata de la búsqueda de un **estimador por intervalos**, un rango de valores entre los que posiblemente se encuentre la cantidad que se estima.

Debemos medir de alguna manera la confianza que podemos tener en el intervalo. Este porcentaje de muestras que dan lugar a intervalos que contienen el auténtico valor del parámetro es el llamado **nivel de confianza**.

Nivel de confianza

El nivel de confianza también se denota por $(1 - \alpha)$ 100% normalmente consideraremos $(1 - \alpha)$, igual a 90%, 95% o 99%.

Así pues, un intervalo de confianza para cierto parámetro con un nivel de confianza de *C*% es un intervalo calculado a partir de una muestra de manera que el procedimiento de cálculo garantiza que el *C*% de las muestras dé lugar a un intervalo que contenga el valor real del parámetro.

La expresión *confianza del 95%* indica confianza en el método utilizado, de manera que el 95% de las veces que apliquemos el método a la misma población obtendremos intervalos que sí contienen el valor del parámetro poblacional.

Intervalo de confianza para la media cuando la población es normal y conocemos la desviación estándar

La variable que queremos estudiar sigue una ley normal de media μ (desconocida) y desviación estándar σ conocida. Disponemos de una muestra aleatoria simple de tamaño n y el valor de la media de la muestra es $\overline{x}$.

Se calculan los intervalos de confianza al nivel de confianza $(1 - \alpha)$% mediante la siguiente expresión:

$$(\text{media de la muestra} - ME, \text{media de la muestra} + ME)$$

donde ME es el **margen de error** que tenemos que calcular, de manera que el $(1 - \alpha)$ % de las muestras produzca un intervalo que contenga el verdadero valor de μ.

El procedimiento que describimos sirve también para variables que no sigan una distribución normal, siempre que la desviación típica sea conocida y que el tamaño de la muestra sea $n > 30$.

Fijamos el nivel de confianza: se acostumbra a considerar $(1 - \alpha)$ igual a 90%, 95% o 99%.

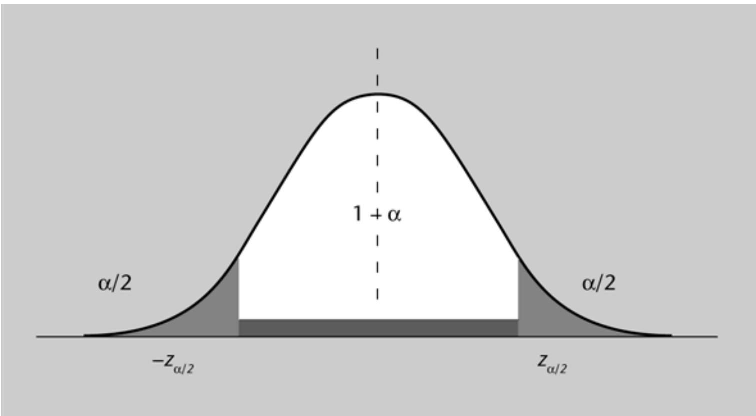
Calculamos el **error estándar** de la media como $\sigma_{\bar{x}} = \sigma / \sqrt{n}$.

Obtenemos el **valor crítico**, que es aquel valor $z_{\alpha/2}$ que hace que:

$$P(Z \geq z_{\alpha/2}) = \alpha/2$$

en el que Z es una variable aleatoria normal $N(0,1)$. Se muestra gráficamente en la figura 11.

Figura 11. Gráfico de intervalo de confianza para μ con desviación típica conocida



Para los niveles de confianza usuales, los valores críticos correspondientes son:

- $(1 - \alpha) = 90\% = 0{,}9$, $\alpha = 0{,}1$ y $z_{\alpha/2} = z_{0,05} = 1{,}645$
- $(1 - \alpha) = 95\% = 0{,}95$, $\alpha = 0{,}05$ y $z_{\alpha/2} = z_{0,025} = 1{,}96$
- $(1 - \alpha) = 99\% = 0{,}99$, $\alpha = 0{,}01$ y $z_{\alpha/2} = z_{0,005} = 2{,}575$

Calculamos el denominado ***margen de error*** (también denominado ***precisión de la estimación***) como $z_{\alpha/2}$ para el error estándar, es decir, como:

$$z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

Nota

Por tanto, el margen de error es la mitad de la longitud del intervalo de confianza.

El intervalo de confianza obtenido con la muestra de partida es:

$$\left(\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right) = \left(\bar{x} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right)$$

o lo que es lo mismo, $\bar{x} \pm ME$.

Es necesario interpretar exactamente los intervalos de confianza. Si se extraen repetida e independientemente muestras aleatorias de n observaciones de la población, entonces el 100(1 – α)% de estos intervalos contendrá el verdadero valor de la media poblacional.

El efecto del tamaño de la muestra

En muchas ocasiones, una vez fijado el nivel de confianza nos marcaremos como objetivo dar el valor del parámetro μ con cierta precisión. La única manera de obtener la precisión deseada consiste en modificar de forma adecuada el tamaño de la muestra. Supongamos que deseamos una precisión o margen de error *ME*; puesto que sabemos que:

Tamaño de la muestra

Es fácil ver que si queremos reducir el ancho del intervalo de confianza a la mitad, deberemos tomar una muestra cuatro veces mayor.

$$ME = z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

Se obtiene el tamaño deseado de la muestra para dicha precisión:

$$n \geq \left(z_{\alpha/2} \right)^2 \frac{\sigma^2}{ME^2}$$

Intervalo de confianza para la media cuando la población es normal y desconocemos la desviación estándar

La variable que queremos estudiar sigue una ley normal de media μ (desconocida) y desviación estándar también desconocida. Disponemos de una muestra aleatoria simple de tamaño n y el valor de la media de la muestra es $\bar{x}$. Entonces:

Calculamos los intervalos de confianza al nivel de confianza (1 – α)%, mediante la siguiente expresión se fija el **nivel de confianza**, que habitualmente se escribe como (1 – α)%.

Calculamos la desviación típica muestral S para obtener el **error estándar** de la media como:

$$s_{\bar{x}} = \frac{s}{\sqrt{n}}$$

Calculamos el **valor crítico**, que es aquel valor $t_{\alpha/2}$ tal que:

$$\mathrm{P}(t_{n-1} \geq t_{n-1,\alpha/2}) = \alpha/2$$

en el que t_{n-1} es una variable aleatoria de Student con $n - 1$ grados de libertad.

Como el **margen de error** es:

$$ME = t_{n-1,\alpha/2}\frac{s}{\sqrt{n}}$$

El intervalo de confianza obtenido con la muestra es el siguiente:

$$\bar{x} \pm ME$$

Intervalo de confianza para la proporción

Interesa conocer la proporción de miembros de la población que poseen una característica específica. Si se toma una muestra aleatoria simple de tamaño *n,* la proporción muestral es un buen estimador de la proporción poblacional. En este apartado se desarrollan intervalos de confianza para la proporción.

Cuando el tamaño de la muestra sea bastante grande, en concreto siempre que el tamaño sea superior a cien, se aplicará el teorema centra del límite, y, como se ha visto en apartados anteriores, la distribución de la proporción muestral sigue una distribución normal estándar $N(0,1)$.

Igual que en los intervalos anteriores se calcula el ***margen de error*** como $z_{\alpha/2}$ multiplicado por el error estándar, es decir:

Nota

El parámetro es p.

El estadístico es $\hat{p}$.

$$ME = z_{\alpha/2}s_{\hat{p}} = z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

El intervalo de confianza obtenido con la muestra de partida será:

$$\hat{p} \pm ME$$

El tamaño de la muestra es $n = \left(z_{\alpha/2}\right)^2 \frac{\hat{p}(1-\hat{p})}{ME^2}$

Ejemplo: un servidor de correo ha recibido 2.000 mensajes, de los cuales 250 son “SPAM”. Construid un intervalo de confianza del 96% para la proporción de mensajes “SPAM”, ¿cuántos correos se han de estudiar en el servidor para poder afirmar que el error entre la proporción de mensajes “SPAM” recibidos y la probabilidad de que el servidor reciba un “SPAM” sea menor que 0,03 con una probabilidad del 95%?

Solución:

El intervalo de confianza del 96% para la proporción de la población se obtiene por medio de la ecuación:

$$\left(\widehat{p} - z_{\alpha/2}\sqrt{\frac{\widehat{p}\cdot(1-\widehat{p})}{n}}, \widehat{p} + z_{\alpha/2}\sqrt{\frac{\widehat{p}\cdot(1-\widehat{p})}{n}}\right)$$

Se deduce que $\widehat{p} = \frac{250}{2000} = 0{,}125,\ n = 2000,\ z_{\alpha/2} = z_{0{,}02} = 2{,}054.$

Por lo tanto, el intervalo de confianza de la proporción poblacional al 96% es

$$\left(0{,}125 - 2{,}054\sqrt{\frac{0{,}125\cdot 0{,}875}{2000}};\quad 0{,}125 + 2{,}054\sqrt{\frac{0{,}125\cdot 0{,}875}{2000}}\right) = (0{,}1098;\ 0{,}1402).$$

Se podría decir que la proporción de todos los mensajes Spam recibidos de la población estarán entre el 10,98% y el 14,02% (con un margen de error del 1,52% al nivel de confianza del 96%).

Se calculará el mínimo tamaño de la muestra necesario para que el error sea menor que 0,03 con una probabilidad del 95% es:

$$n \geq (z_{\alpha/2})^2\,\frac{\widehat{p}\cdot(1-\widehat{p})}{ME^2} = (z_{0{,}025})^2\,\frac{0{,}125\cdot 0{,}875}{0{,}03^2} = 1{,}96^2\cdot\frac{0{,}109}{0{,}0009} = 466{,}75$$

Por tanto, se deben estudiar 467 mensajes.

Ejemplo con Minitab: en el ejemplo anterior se comparan los intervalos de confianza al 90 y el 99%, manteniendo constante el tamaño de la muestra, para contestar a la siguiente pregunta: Conforme aumenta la amplitud de un intervalo de confianza, ¿aumenta o disminuye el nivel de confianza asociado? En las figuras 12 y 13 utilizamos Minitab para analizar ambos escenarios.

Figura 12. Resultado del Intervalo de confianza del 90% con Minitab

```
Test and CI for One Proportion

Test of p = 0,125 vs p not = 0,125


Sample    X     N  Sample p          90% CI           Z-Value  P-Value
1       250  2000  0,125000  (0,112836; 0,137164)     0,00     1,000

Using the normal approximation.
```

Figura 13. Resultado del Intervalo de confianza del 99% con Minitab

```
Test and CI for One Proportion

Test of p = 0,125 vs p not = 0,125


Sample    X     N  Sample p          99% CI           Z-Value  P-Value
1       250  2000  0,125000  (0,105951; 0,144049)     0,00     1,000

Using the normal approximation.
```

Notar que al aumentar el nivel de confianza, deberemos ampliar la amplitud del intervalo a fin de “abarcar” un rango mayor para el parámetro poblacional estimado.

Intervalo de confianza para la varianza

¿Cómo se puede construir un intervalo de confianza para la varianza poblacional?

Primero se fijará el nivel de confianza $1 - \alpha$. Se calculará el estadístico.

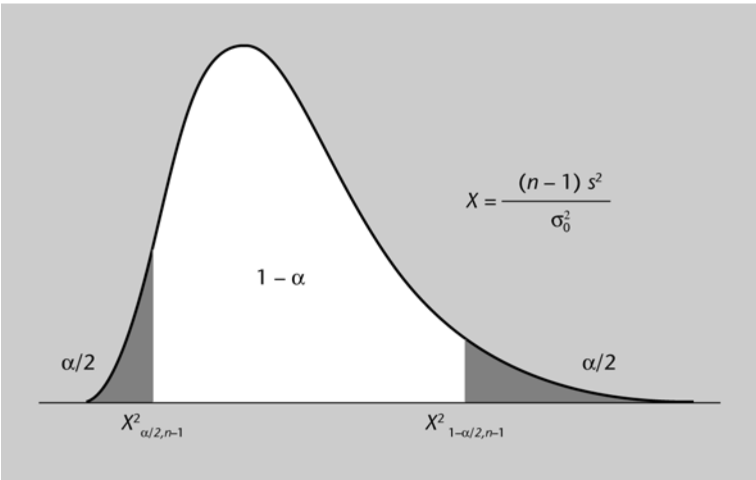
$$\chi^2 = \frac{(n-1)s^2}{\sigma^2}$$

es una observación de una variable aleatoria χ^2 con $n - 1$ grados de libertad.

Donde s^2 es la varianza muestral de una muestra aleatoria de tamaño n tomada de una población normal de varianza σ^2.

La figura 14 muestra los valores de la distribución χ^2_{n-1} que cortan una probabilidad de $\alpha/2$ en las dos colas, es decir, los puntos críticos $\chi^2_{n-1,\alpha/2}$ y $\chi^2_{n-1,\,1-\alpha/2}$.

Figura 14. Gráfico de intervalo de confianza de la varianza



Ejemplo de intervalo de confianza para la varianza

Una empresa de autobuses urbanos espera que las horas de llegada en diversas paradas tengan poca variabilidad. La varianza de la muestra de 10 tiempos de llegada de autobús fue s^2 = 4,8 minutos^2. Suponiendo que la población de tiempos de llegada tiene una distribución normal, se desea determinar un intervalo de confianza del 95% para la varianza poblacional de los tiempos de llegada.

El estadístico de prueba: $\chi^2 = \frac{(n-1)s^2}{\sigma^2}$ tiene una distribución chi-cuadrado con $n - 1 = 9$ grados de libertad. Determinamos los valores $\chi^2_{9,0,975} = 16{,}0471$ y $\chi^2_{9,0,025} = 45{,}7222$.

El intervalo de confianza para la varianza de la población será:

$$\left[\frac{(n-1)s^2}{\chi^2_{n-1,\,1-\alpha/2}};\frac{(n-1)s^2}{\chi^2_{n-1,\alpha/2}}\right] = \left[\frac{9\cdot 4,8}{45,7222};\frac{9\cdot 4,8}{16,0471}\right] = [0,94;2,69]\text{minutos}$$

La raíz cuadrada de esos valores será el intervalo de confianza de 95% para la desviación estándar: $0,97 \leq \sigma \leq 1,64$.

6. Contrastes de hipótesis para una población

En este apartado se desarrollan métodos para contrastar hipótesis que permiten comparar la validez de una conjetura o afirmación utilizando datos muestrales. El proceso comienza cuando un investigador formula una hipótesis sobre la naturaleza de una población. La formulación de esta hipótesis implica la elección entre dos opciones; a continuación, el investigador selecciona una opción basándose en los resultados de un estadístico calculado a partir de una muestra aleatoria de datos.

He aquí algunos ejemplos de problemas representativos:

1) Un investigador quiere saber si una propuesta de reforma fiscal es acogida de igual forma por hombres y mujeres. Para analizar si es así, recoge las opiniones de una muestra aleatoria de hombres y mujeres.

2) Una compañía recibe un cargamento de piezas. Sólo puede aceptar el envío si no hay más de un 5% de piezas defectuosas. La decisión de si aceptar la remesa puede basarse en el examen de una muestra aleatoria de piezas.

3) Una profesora está interesada en valorar la utilidad de hacer controles regularmente en un curso de estadística. El curso consta de dos partes y la profesora realiza estos controles sólo en una de ellas. Cuando acaba el curso, compara los conocimientos de los estudiantes en las dos partes del curso mediante un examen final y analiza la hipótesis de que los controles aumentan el nivel medio de conocimientos.

Los ejemplos propuestos tienen algo en común. La hipótesis se formula sobre la población y las conclusiones sobre la validez de esta hipótesis se basan en la información muestral. El test o contraste será la herramienta que nos permitirá extraer conclusiones a partir de la diferencia entre las observaciones y los resultados que se deberían obtener si la hipótesis de partida fuese cierta.

Planteamiento del contraste de hipótesis

En la prueba de hipótesis se comienza proponiendo una hipótesis de partida acerca de un parámetro poblacional. Esta hipótesis se llama **hipótesis nula** y se representa como $\boldsymbol{H_0}$. A continuación se define otra hipótesis, la **hipótesis alternativa**, que es la opuesta de lo que se afirma en la hipótesis nula. La hipótesis alternativa se representa como $\boldsymbol{H_1}$. El procedimiento para probar una hipótesis comprende el uso de datos de una muestra para probar las dos aseveraciones representadas por H_0 y H_1.

Hipótesis

Con la misma hipótesis nula podemos estudiar varias hipótesis alternativas.

Las hipótesis expresan una afirmación sobre el valor del parámetro. Podemos tener una hipótesis nula del tipo $\boldsymbol{H_0}$: $\theta = \theta_0$.

La hipótesis alternativa puede ser unilateral, como H_1: $\theta > \theta_0$ o H_1: $\theta < \theta_0$, o bilateral, como H_1: $\theta \neq \theta_0$.

Una vez planteadas las hipótesis nula y alternativa, debemos tomar una decisión a partir de las observaciones. Por otro lado, existen dos decisiones posibles:

1) Aceptar la hipótesis nula.

2) Rechazar la hipótesis nula.

Errores en el contraste

Con el fin de llegar a una de estas dos conclusiones, se adopta una **regla de decisión** basada en la evidencia muestral. Por consiguiente, *no se puede saber con seguridad* si la hipótesis nula es cierta o falsa. Por tanto, cualquier regla de decisión adoptada tiene cierta probabilidad de llegar a una conclusión falsa. Como se indica en la tabla 1, pueden cometerse dos tipos de errores. Un error que se puede cometer, llamado **error de tipo I**, es rechazar una hipótesis nula cierta. Si la regla de decisión es tal que la probabilidad de rechazar la hipótesis nula cuando es cierta es α, entonces α se llama **nivel de significación** del contraste. La probabilidad de aceptar la hipótesis nula cuando es cierta es $(1 - \alpha)$. El otro error posible, llamado **error de tipo II**, ocurre cuando se acepta una hipótesis nula falsa. La probabilidad de cometer este tipo de error, cuando la hipótesis nula es falsa, se denota por β. Entonces, la probabilidad de rechazar una hipótesis nula falsa es $(1 - \beta)$, y se denomina **potencia del contraste.**

Regla de decisión

Error de tipo I: rechazar una hipótesis nula cierta.

Error de tipo II: aceptar una hipótesis nula falsa.

Nivel de significación: la probabilidad de rechazar una hipótesis nula que es cierta (esta probabilidad a veces se expresa en %, con lo que nos referimos a un contraste de significación α como un contraste al nivel 100 α%).

Potencia: la probabilidad de rechazar una hipótesis nula que es falsa.

Tabla 1. Errores y decisiones correctas en contrastes de hipótesis

		Condición de la población	
		H_0 verdadera	**H_0 falsa**
Decisión	**Aceptar H_0**	Decisión correcta	Error de tipo II
	Rechazar H_0	Error de tipo I	Decisión correcta

Atención

Un nivel α **= 0,05** significa que aunque la hipótesis nula sea cierta, los datos de cinco de cada cien muestras nos la harán rechazar. Es decir, aceptamos que podemos rechazar la hipótesis nula equivocadamente cinco de cada cien veces.

Para plantear y resolver un contraste de hipótesis, es necesario:

1) Fijar las hipótesis nula y alternativa.

2) Fijar un nivel de significación.

3) Determinar el estadístico de contraste y su ley.

4) A partir de aquí, tenemos dos métodos posibles:

4a) Calcular el *p*-valor asociado a nuestro estadístico de contraste calculado. Comparar el *p*-valor con el nivel de significación y tomar una decisión.

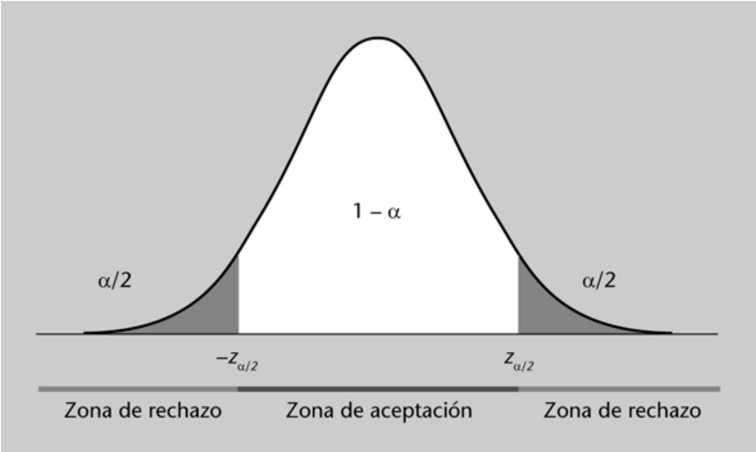
4b) Calcular el valor crítico. Comparar el valor crítico con el estadístico de contraste y tomar una decisión.

Zona de aceptación y zona de rechazo de la hipótesis nula

Ejemplo 1. "Contraste bilateral"

La parte del gráfico (figura 15) sombreada en rojo corresponde a la zona en la que rechazamos la hipótesis nula. La zona sin sombrear corresponde a la región de aceptación de la hipótesis nula.

Figura 15. Gráfico que muestra la zona de aceptación y de rechazo de la hipótesis nula en un contraste bilateral



Recordad

Si tenemos una muestra de tamaño n de una distribución $N(\mu, \sigma^2)$, entonces

$$Z = \frac{\overline{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

sigue una distribución normal estándar.

Para determinar el valor $z_{\alpha/2}$, sólo hay que imponer que el error de tipo I (probabilidad de rechazar H_0 cuando es cierta) sea menor o igual que el nivel de significación α. Por ejemplo, para α = 0,05 encontramos (por ejemplo, en las tablas de la normal) que $z_{\alpha/2}$ = 1,96.

Para decidir si rechazamos la hipótesis nula o no, usaremos el llamado **estadístico de contraste.** Un estadístico de contraste es una función de la muestra cuya distribución conocemos bajo la hipótesis nula.

Validez del método

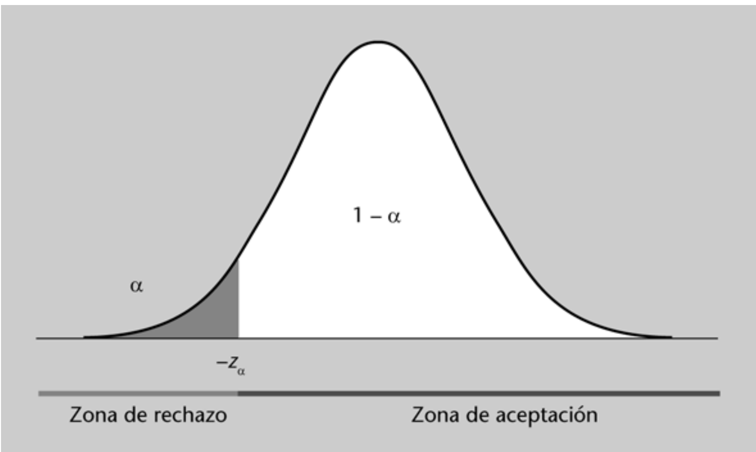
El método es el mismo para cualquier distribución simétrica, así que también sirve si el estadístico de contraste sigue una distribución *t* de Student.

- Aceptaremos H_0 si $|z| \leq z_{\alpha/2}$
- Rechazaremos H_0 si $|z| \geq z_{\alpha/2}$

Ejemplo 2. "Contraste unilateral inferior"

La parte del gráfico (figura 16) sombreada corresponde a la zona de rechazo de la hipótesis nula. La zona sin sombrear corresponde a la región de aceptación de la hipótesis nula.

Figura 16. Gráfico que muestra la zona de rechazo de la hipótesis nula en un contraste unilateral inferior



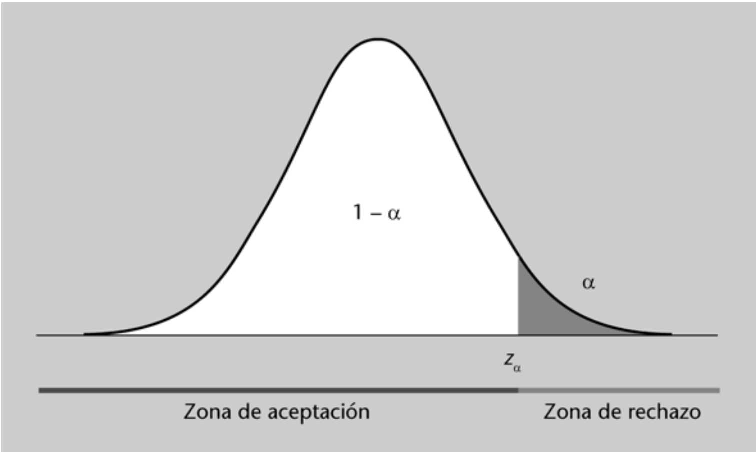
Para α= 0,05 encontramos que $-z_\alpha = -1{,}65$. En este contraste unilateral se dice que la probabilidad de la cola de la izquierda debe ser α.

- Aceptaremos H_0 si $\boldsymbol{Z} \geq -z_\alpha$
- Rechazaremos H_0 si $\boldsymbol{Z} < -z_\alpha$

Ejemplo 3. "Contraste unilateral superior"

La parte del gráfico (figura 17) sombreada en rojo corresponde a la zona en la que rechazamos la hipótesis nula. La zona sin sombrear corresponde a la región de aceptación de la hipótesis nula.

Figura 17. Gráfico que muestra la aceptación o no de la hipótesis nula en un contraste unilateral superior



Para $\alpha = 0{,}05$ encontramos que $z_\alpha = 1{,}65$. En este contraste unilateral se dice que la probabilidad de la cola de la derecha debe ser α.

- Aceptaremos H_0 si $\mathbf{Z} \leq z_\alpha$
- Rechazaremos H_0 si $\mathbf{Z} > z_\alpha$

El *p*-valor

Existe otro método para examinar el contraste de la hipótesis nula. Obsérvese que si se utiliza un nivel de significación bajo se reduce la probabilidad de rechazar una hipótesis nula verdadera. Eso modificaría la regla de decisión para que fuera menos probable que se rechazara la hipótesis nula, independientemente de que fuera verdadera o no. Evidentemente, cuanto menor es el nivel de significación al que se rechaza una hipótesis nula mayores son las dudas sobre su veracidad. En lugar de contrastar hipótesis a los niveles preasignados de significación, los investigadores a menudo hallan el nivel menor de significación al que se puede rechazar una hipótesis nula.

El *p*-valor es el menor nivel de significación al que puede rechazarse una hipótesis nula.

El criterio del *p*-valor es: rechazar H_0 si el *p*-valor < α.

Interpretación del *p*-valor

Se considera una muestra aleatoria de *n* observaciones procedentes de una población que sigue una distribución normal de media μ y desviación estándar σ y la media muestral calculada $\bar{x}$. Se ha contrastado la hipótesis nula $H_0 : \mu = \mu_0$ frente a la alternativa $H_1 : \mu > \mu_0$

El *p*-valor del contraste es:

$$p\text{-valor} = P\left(\frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}} \geq z_p \middle| H_0 : \mu = \mu_0 \right)$$

donde z_p es el valor normal estándar correspondiente al menor valor de significación al que puede rechazarse la hipótesis nula. La mayoría de los programas informáticos estadísticos calculan el *p*-valor, este suministra más información sobre el contraste basándose en la media muestral observada, por lo que se utiliza frecuentemente en muchas aplicaciones estadísticas.

Ejemplo de aplicación del *p*-valor: un grupo editorial emite un periódico especializado en información económica. El director del periódico desea saber si el número medio de ejemplares diarios producidos y no vendidos es menor de 400. Para dar respuesta a esta pregunta, se toma una muestra formada por los resultados correspondientes a 172 días elegidos de forma aleatoria. La media de dicha muestra es de 407 ejemplares no vendidos, con una desviación estándar de 38.

Utilizando un nivel de significación de 0,05, realizad un contraste de hipótesis para responder razonadamente a la pregunta del director del periódico.

Solución:

1) Si se hace el contraste H_0: media poblacional = 400 contra H_1: media poblacional ≠ 400.

Primero se calcula el estadístico de contraste para decidir si rechazamos la hipótesis nula o no.

La desviación estándar de la muestra es: $\frac{S}{\sqrt{n}} = \frac{38}{\sqrt{172}} = 2{,}89$.

El estadístico será $z = \frac{407 - 400}{2{,}89} = 2{,}42$, este valor es una observación de una distribución *N*(0,1).

En este caso, por ser un contraste bilateral se divide el nivel de significación α por igual entre las dos colas de la distribución normal. Por lo tanto, la probabilidad de que *Z* sea superior $z_{\alpha/2}$ o inferior a $-z_{\alpha/2}$ es α. En este caso, el

p-valor es la suma de las probabilidades de la cola superior y la cola inferior. El p-valor correspondiente al contraste de dos colas es:

$$p - valor = 2P\left(\left|\frac{\overline{x}-\mu_0}{\sigma/\sqrt{n}}\right| > z_{\alpha/2}\right);$$

$$P(Z > |2,42|) = P(Z > 2,42) + P(Z < -2,42) = 2 \cdot 0,00776 = 0,01552$$

Como 0,01552 es menor que el nivel de significación propuesto (α = 0,05), se rechaza la hipótesis nula. No se puede afirmar que el número medio de ejemplares diarios producidos y no vendidos sea de 400. Se acepta que es distinto de 400.

2) Si se hace el contraste H_0: media poblacional = 400 contra H_1: media poblacional > 400, entonces el p-valor es la probabilidad "es la cola de la derecha":

$$p - valor = P(Z) > z_\alpha$$

$P(Z > 2,42) = 0,00776 < \alpha \Rightarrow$ Se rechaza la hipótesis nula.

Se acepta la hipótesis alternativa, por lo tanto, se acepta que el número medio de ejemplares diarios producidos y no vendidos es mayor de 400.

3) Si se realiza el contraste H_0: media poblacional = 400 contra H_1: media poblacional < 400, entonces el p-valor es la probabilidad "es la cola de la izquierda":

$$p - valor = P(Z) < z_\alpha$$

$P(Z < 2,42) = 1 - 0,00776 = 0,99224 > \alpha \Rightarrow$ No se puede rechazar la hipótesis nula.

Se rechazará la hipótesis alternativa, luego el número medio de ejemplares diarios producidos y no vendidos no es menor de 400.

Por tanto, a la vista de los resultados de los tres contrastes, la contestación a la pregunta del director sería:

"El número medio de ejemplares diarios producidos y no vendidos es mayor de 400".

Para calcular el p-valor se suele utilizar un software estadístico, como se verá en ejemplos resueltos con Minitab.

Otro procedimiento: para resolver contrastes bilaterales utilizando intervalos de confianza.

Ejemplo: supongamos que se plantea el siguiente contraste bilateral:

$$H_0: \mu = 280,\ H_1: \mu \neq 280$$

Para probar esta hipótesis con un nivel de significación $\alpha = 0{,}05$, el tamaño de la muestra es 36 y se determinó que la media muestral $\overline{x} = 278{,}5$ y la desviación estándar de las muestras $s = 12$. Sustituyendo estos resultados con $z_{0,025} = 1{,}96$, vemos que el intervalo de confianza del 95% para la media de la población es:

$$\overline{x} \pm 1{,}96\frac{s}{\sqrt{n}} \;\; ; \;\; 278{,}5x \pm 1{,}96\frac{12}{\sqrt{36}} \;\; ; 278{,}5 \pm 3{,}92$$

El intervalo será: (274,58; 282,42).

El resultado permite llegar a la conclusión de que, con un 95% de confianza, la media para la población está entre 274,58 y 282,42. Como el valor supuesto de la media de la población μ_0= 280 está en el intervalo de confianza, la conclusión del contraste es que no se puede rechazar la hipótesis nula, por tanto, aceptamos la hipótesis de que: H_0: μ =280.

Ejemplo de inferencia para una población (utilizando Minitab)

Una característica importante en el diseño de una página web es el tiempo que el usuario tardará en abrir la página, que se considera una variable normal. Con el objetivo de estimar el tiempo medio, se seleccionan al azar 101 páginas, entre las que ha diseñado una empresa el último año, obteniéndose los datos siguientes (en centésimas de segundo):

Tabla 2. Tiempo de descarga de páginas web

Tiempo de descarga	55	60	62	64	65	69
Número de páginas	11	21	26	19	15	9

Observación: se crea un fichero de datos en la hoja de Minitab, introduciendo los datos de forma unitaria.

a) Se comprueba que la colección de datos sigue una distribución aproximadamente normal.

b) Puede considerarse que el tiempo medio de apertura de las páginas de esta empresa es de 62 centésimas de segundo, con un nivel de confianza del 90%. ¿Qué resultado se obtiene? Razónese la respuesta del contraste a través del *p*-valor.

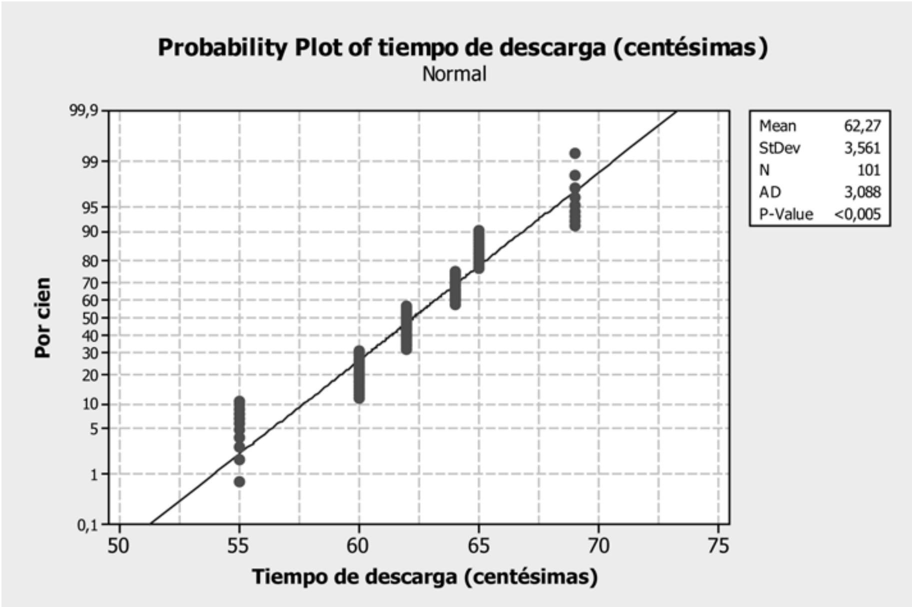
c) Calcúlese un intervalo de confianza a nivel del 90% para el tiempo medio y coméntese si el resultado obtenido es coherente con el resultado esperado.

d) Finalmente, se realizará el mismo contraste que en el apartado b), pero suponiendo esta vez que no se conoce la desviación estándar.

Solución:

a) Para comprobar la normalidad de los datos, se selecciona ***Stat*** > ***Basic Statistics*** > ***Normality Test***. Así se obtiene el gráfico de la figura 18.

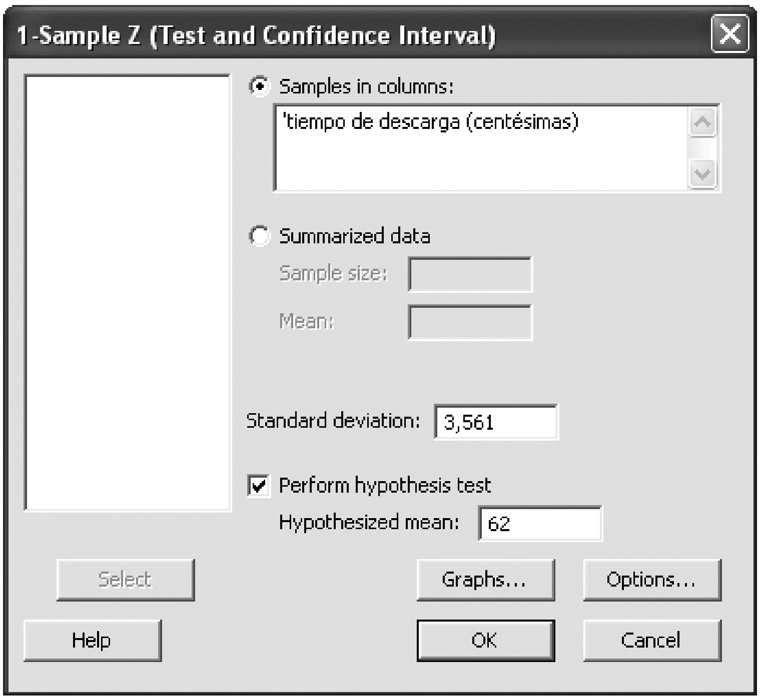
Figura 18. Gráfico de normalidad



Observando el *p*-valor se puede concluir que los datos siguen una distribución normal. Pudiendo asegurar que X sigue una distribución normal, la media muestral también sigue una distribución normal.

b) El contraste de hipótesis será H_0: $\mu = 62$ vs. H_1: $\mu \neq 62$. Es un contraste bilateral a un nivel de confianza de 0,90 (figura 19).

Figura 19. Pasos a seguir para realizar el contraste de hipótesis



Los resultados de Minitab son los que muestra la figura 20.

Figura 20. Resultados del contraste de hipótesis e intervalo de confianza del 90% (desviación típica población conocida)

```
One-Sample Z: tiempo de descarga (centésimas)

Test of mu = 62 vs not = 62
The assumed standard deviation = 3,561

Variable  N    Mean  StDev  SE Mean       90% CI           Z
tiempo  101  62,267  3,561    0,354  (61,685; 62,850)  0,75

Variable                        P
tiempo de descarga (cent.)  0,451
```

Se observa que el *p*-valor es 0,451, por lo tanto, como *p*-valor > α = 0,10, no se puede rechazar la hipótesis nula, luego se acepta que el tiempo medio es de 62 centésimas por segundo.

c) El intervalo de confianza para el tiempo medio es (**61,685**; **62,850**), es coherente con los resultados esperados, ya que contiene al valor medio de 62 centésimas de segundo.

d) Análogamente se realiza el contraste de hipótesis para la media de la población con desviación típica desconocida, se selecciona ***Stat*** > ***Basic Statistic*** > ***1-Sample t***, obteniéndose los resultados de la figura 21.

Figura 21. Resultados del contraste de hipótesis e intervalo de confianza del 90% (desviación típica población desconocida)

```
One-Sample T: tiempo de descarga (centésimas)

Test of mu = 62 vs not = 62

Variable  N    Mean  StDev  SE Mean       90% CI           T
tiempo    101  62,267  3,561    0,354  (61,679; 62,856)  0,75

Variable                        P
tiempo de descarga (cent.)    0,452
```

El *p*-valor es 0,452 > 0,10, nos indica que se puede aceptar la hipótesis de que el tiempo medio es de 62 centésimas por segundo.

Continuando con el mismo ejemplo, se va a considerar que una página no es satisfactoria cuando tarde en ser descargada más de 68 centésimas. Los programadores afirman que el porcentaje de páginas para las que el tiempo de descarga no es satisfactorio no supera el 10%.

e) Se calculará un intervalo de confianza para la proporción de páginas no satisfactorias, a un nivel de confianza del 95%.

f) ¿Hay evidencias, al nivel 0,05, para rechazar la afirmación de los programadores? Se plantearán las hipótesis que se deben contrastar y se efectuará el contraste.

e) Para calcular el intervalo de confianza de la proporción de páginas no satisfactorias, a un nivel de confianza del 95%, se selecciona ***Stat*** > ***Basic Statistics >1 Proportion*** (figura 22).

Observando la figura 23 de datos, se ve que únicamente hay 9 páginas que superan las 68 centésimas de segundo, o lo que es lo mismo, 9 páginas de las 101 se considera el tiempo de descarga no satisfactorio.

Figura 22. Pasos a seguir para obtener un intervalo de confianza del 95% para la proporción

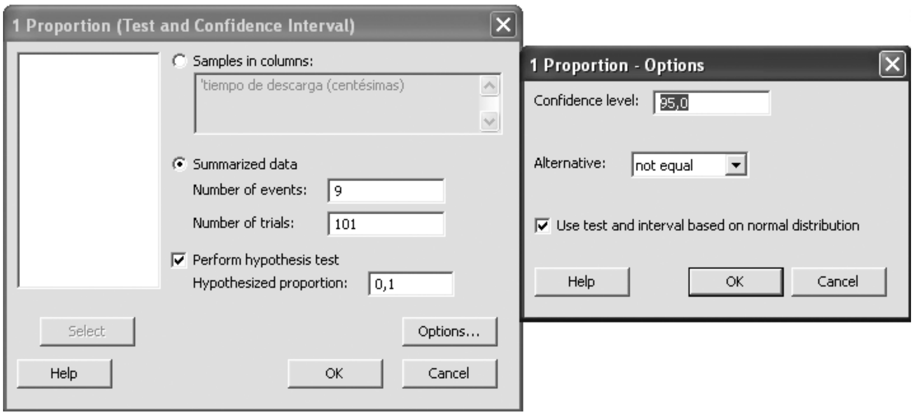


Figura 23. Resultados del intervalo de confianza del 95% para la proporción de páginas no satisfactorias

```
Test and CI for One Proportion

Test of p = 0,1 vs p not = 0,1

Sample X   N   Sample p         95% CI          Z-Value  P-Value
  1    9  101  0,089109  (0,033546; 0,144671)   -0,36    0,715

Using the normal approximation.
```

El intervalo de confianza obtenido con un nivel de confianza del 95% es **(0,033546; 0,144671)**.

f) Debemos plantear un contraste unilateral para la proporción de páginas no satisfactorias:

$$\left.\begin{array}{l} H_0 : p = 0{,}1, \\ H_1 : p > 0{,}1, \end{array}\right\},$$ donde p representa la proporción de páginas para las que el tiempo de descarga no es satisfactorio (figura 24).

Figura 24. Pasos a seguir para realizar el contraste de hipótesis

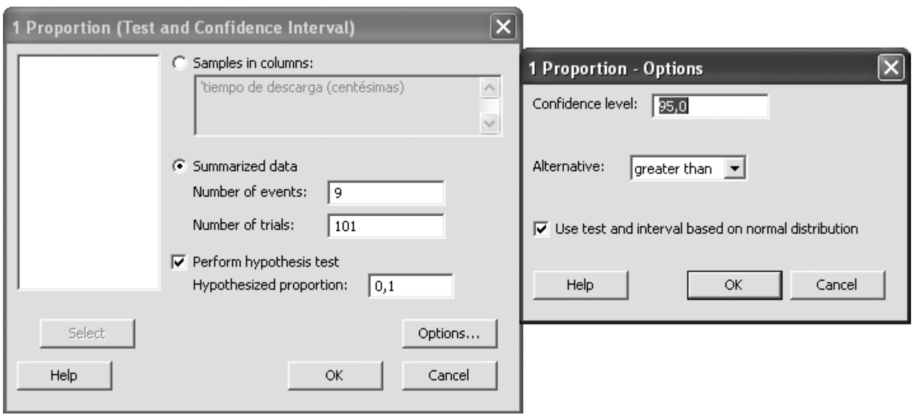


Figura 25. Resultados del contraste de hipótesis para la proporción de páginas

Test and CI for One Proportion

```
Test of p = 0,1 vs p > 0,1


                            95% Lower
Sample  X    N  Sample p      Bound  Z-Value  P-Value
1       9  101  0,089109   0,042479    -0,36    0,642

Using the normal approximation.
```

Según se muestra en la figura 25, el *p*-valor del contraste vale lo siguiente: *p*-valor = 0,642. Como es mayor que 0,05, se acepta la hipótesis nula, luego se acepta la afirmación de los programadores de que el porcentaje de páginas no supera el 10%.

Resumen

En este módulo se presentan las distribuciones muestrales. Se analiza cómo seleccionar una muestra aleatoria simple, cómo se pueden emplear los datos obtenidos con ella para desarrollar estimaciones puntuales de los parámetros de población. La distribución de probabilidad de estas variables aleatorias se llama *distribución muestral*. En particular, se describen las distribuciones de la media de la muestra $\bar{x}$, de la proporción muestral $\hat{p}$ y de la varianza muestral s^2. Después de desarrollar las fórmulas de la desviación típica o error estándar para esos estimadores, se indica que el teorema central del límite es la base para usar una distribución normal de probabilidades y aproximar esas distribuciones muestrales en el caso de muestra grande.

Además, también se desarrollan estimaciones de intervalos de confianza de parámetros de una población. En este módulo se han utilizado la distribución Z normal estándar, la t de Student y la chi-cuadrado χ^2 para construir intervalos de confianza. Se determina el tamaño de muestra necesario para que los estimadores de intervalo de μ y de p tengan un nivel especificado de precisión.

Finalmente, en este módulo se ha presentado la metodología para realizar contrastes clásicos de hipótesis, comenzando con los argumentos para tomar decisiones en condiciones de incertidumbre. Las decisiones se toman rechazando una hipótesis nula si hay pruebas contundentes a favor de la hipótesis alternativa. Pueden cometerse dos tipos de error: un error de tipo I, que se comete cuando se rechaza la hipótesis nula, cuando es verdadera, y un error de tipo II, que se comete cuando no se rechaza la hipótesis nula, cuando no es verdadera, presentando diversos métodos y reglas de decisión específicos para realizar contrastes. La regla de rechazo para todos los procedimientos implica comparar el valor del estadístico con un valor crítico y también utilizando el p-valor para pruebas de hipótesis, la regla es rechazar la hipótesis nula siempre que el p-valor sea menor que α.

Ejercicios de autoevaluación

1) Una biblioteca presta un promedio de μ =320 libros por día, con desviación estándar σ = 75 libros. Se tiene una muestra de 30 días de funcionamiento, y $\bar{x}$ es la cantidad de la media de la muestra de libros prestados en un día.

a) Presente la distribución muestral de $\bar{x}$.
b) ¿Cuál es la distribución estándar de $\bar{x}$?
c) ¿Cuál es la probabilidad de que la media de una muestra de 30 días sea entre 300 y 400 libros?
d) ¿Cuál es la probabilidad de que la media de una muestra sea de 325 o más prestamos diariamente?

2) Un investigador informa los resultados de una encuesta diciendo que el error estándar de la media es de 20.La desviación estándar de la población es de 500.

a) ¿De qué tamaño fue la muestra que se usó en esta encuesta?
b) ¿Cuál es la probabilidad de que el error estimado quede a ± 25 o menos de la media de la población?

3) Cada curso escolar, una prestigiosa universidad oferta becas a sus estudiantes para ampliar estudios en el extranjero. De la experiencia recogida en anteriores convocatorias, se observa que las calificaciones medias de los expedientes aspirantes a obtener una beca se distribuyen según una normal de media 6,9 puntos y desviación estándar 0,7 puntos. Para entender la aplicación del teorema central del límite, generar con Minitab 50 muestras aleatorias de 100 observaciones cada una, que corresponden a la población normal anterior $N(6{,}9,\ 0{,}7)$.

a) Calcular en una nueva columna la media de las 50 muestras anteriores.
b) Comentar los resultados haciendo referencia al teorema central del límite.
c) Realiza el *dotplot* asociado a una de las muestras.
d) Compara estos resultados con la media de la población, y el valor de la desviación estándar de la media muestral con la desviación estándar de la población y explica la relación entre ambos valores.

4) Un estudio previo nos dice que el servicio de préstamo diario de libros de las bibliotecas de una ciudad sigue una distribución normal con una media de 300 ejemplares prestados, con una desviación estándar de 10. Una inspección quiere verificar si estos datos son correctos. Para hacerlo, coge una muestra de los préstamos diarios de 10 bibliotecas y obtiene una media de 285 ejemplares prestados.

a) ¿Cuál es la probabilidad de que si la media es verdaderamente de 300 ejemplares prestados se obtenga una media de préstamos igual o inferior a los 285 ejemplares en las 10 bibliotecas que componen la muestra?
b) Determinar un intervalo de confianza del 90% para la media de préstamos teniendo en cuenta los datos de la muestra.
c) ¿Qué decisión lógica debería tomar el inspector?

5) En la página web de una editorial aparecen dos números de teléfono. Hemos comprobado, después de analizar 400 llamadas del teléfono, que el intervalo entre llamadas tiene una varianza de 2.

Suponiendo normalidad, indicad si podemos considerar, a un nivel de confianza del 90%, que la varianza del intervalo entre llamadas del primer número es inferior a 1,7.

6) El responsable de comunicaciones de un centro de documentación afirma que la media del tiempo de transferencia de un fichero de tamaño 2Mb es superior a 30 segundos. Para comprobar esta afirmación se tomó una muestra de tiempos de transferencia de 12 ficheros de 2Mb, obteniendo que la media y la desviación estándar muestrales valen $\bar{x} = 30{,}2,\ s = 1{,}833$ (en segundos).

a) Suponiendo que el tiempo de transferencia se distribuye normalmente a partir de los datos muestrales obtenidos, ¿tenemos suficientes evidencias para aceptar la afirmación del responsable? (Tomad α = 0,05). Encontrad el *p*-valor del contraste.

Si además de disponer de estas observaciones nos hubiesen dado como información adicional (obtenida de experiencias previas) que la varianza del tiempo de transferencia es de $\sigma^2 = 9{,}2$ segundos^2, ¿hubiéramos llegado a la misma conclusión que en el apartado anterior? Encontrad el *p*-valor del contraste (Tomad α = 0,05).

Solucionario

1)

a) Normal con $\mu = 320$ y desviación típica 13,69
b) 13,69
c) 0,8558
d) 0,3557

2)

a) 625
b) 0,7888

3)

De esta manera obtenemos las 50 muestras con 100 observaciones cada una.

Worksheet 1 ***

↓	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15	C16
1	6,02255	7,51039	7,56941	6,73114	6,89504	7,27531	6,49352	7,07449	7,72782	8,01131	6,51824	7,31135	5,70901	7,18919	6,23128	7,986
2	5,82796	7,23382	8,63014	7,67923	6,82915	7,50193	6,80251	6,83211	6,08951	7,08834	7,44647	7,36977	6,94125	6,82865	8,08839	6,073
3	5,89460	6,19736	7,48944	7,16782	7,30042	6,23994	5,55889	5,92865	6,65836	7,20623	6,29604	6,50873	7,42008	6,63113	7,64693	7,277
4	7,91373	6,63874	6,88954	7,74717	7,75320	6,93165	6,16663	5,94975	7,84482	7,35052	6,62743	7,61428	6,31124	6,59273	6,66213	5,852
5	6,96531	6,54096	7,83251	6,71421	5,76998	6,48673	6,47134	6,87821	7,57085	6,96200	7,26595	7,65586	7,88420	8,22074	7,28688	5,476
6	6,38379	8,14155	5,26726	7,28067	6,89716	6,26790	7,12585	5,82249	6,81010	7,49986	6,99193	5,76598	7,32813	5,52918	7,05477	5,853
7	6,98679	7,50650	7,06138	7,30100	6,61602	7,20820	8,00421	6,49912	8,02260	7,28351	5,36631	6,93591	7,84171	6,61069	6,50505	6,263
8	6,28190	8,55299	8,20722	7,29348	6,51880	7,74509	6,95879	7,46706	7,55387	7,82200	7,22971	7,12365	7,03669	6,01653	8,28156	6,957
9	5,85308	6,42670	7,24762	7,50398	6,46682	7,32013	6,42494	5,69820	6,13376	6,79857	7,15053	6,40466	6,38444	6,24852	6,05964	6,250

Project Man...

Current Worksheet: Worksheet 1 18:19

a) En la columna C101 se muestran las medias maestrales.

Worksheet 1 ***

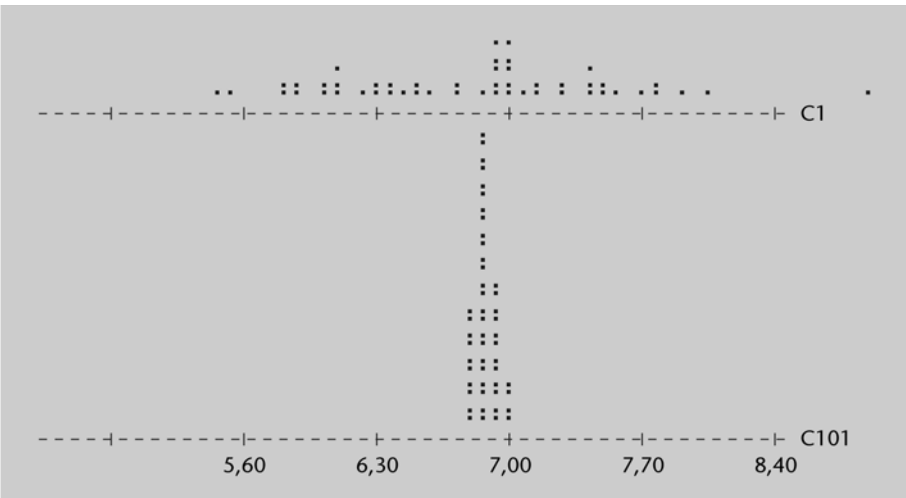
↓	C98	C99	C100	C101
1	5,49495	7,33352	6,49861	6,84032
2	7,27693	5,57569	5,13667	6,90402
3	6,86654	6,96175	7,56928	6,87066
4	7,85956	6,10293	7,09721	6,87309
5	7,68902	4,92515	7,13723	6,89323
6	6,51449	5,78576	7,18556	6,81035
7	7,46093	6,67470	5,89898	6,87570
8	5,41243	7,05719	7,60637	6,98587
9	8,66592	7,03988	7,55059	6,85254

Project Man...

Current Worksheet: Worksheet 1

b)

***dotplot*: C1; C101**



c) Tras haber generado 50 muestras de datos provenientes de una distribución normal de media 6,9 y desviación estándar 0,7, observamos que el primer *dotplot* parece corresponder a una distribución normal.

Asimismo, el segundo *dotplot* corresponde a la distribución de las medias de las muestras y también corresponde a una distribución normal.

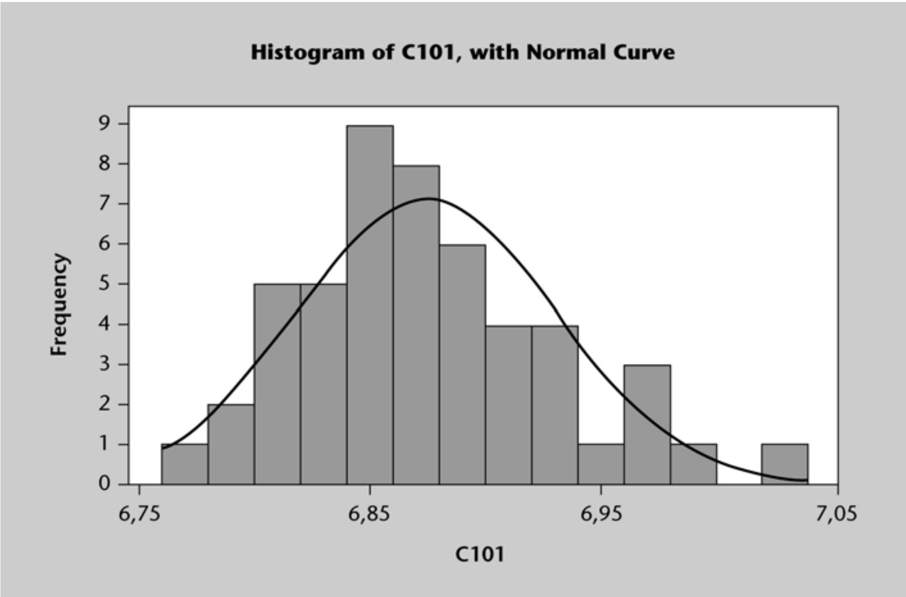
Esto indica que las medias de estas muestras siguen una distribución normal. Esta propiedad es la que enuncia el TCL, sea cual sea la distribución de los datos, la media muestral (con un tamaño de muestra *n* suficientemente grande) de una colección de datos sigue una distribución normal.

d) Estudiaremos la distribución de estas medias muestrales:

Descriptive Statistics: C101

Variable	N	Mean	Median	TrMean	StDev	SE Mean
C101	50	6,9035	6,8914	6,9016	0,0660	0,0093

Variable	Minimum	Maximum	Q1	Q3
C101	6,7837	7,0412	6,8507	6,9603



El histograma de frecuencias se aproxima a la curva normal, es simétrica.

La media muestral coincide con la media de la población, $\mu = \overline{x} = 6{,}9$.

La desviación estándar de la media muestral será aproximadamente el error estándar.

Si la variable tiene desviación estándar conocida *s* (en la población), el error estándar se puede calcular como:

$$\frac{\sigma}{\sqrt{n}}$$

Como consecuencia, podemos decir que la media muestral sigue una distribución normal $N(\mu, \frac{\sigma}{\sqrt{n}})$, que se puede aproximar a una $N(0,1)$, realizando un cambio de variable (tipificación): $Z = \frac{X - \mu}{\sigma/\sqrt{n}}$.

4)

a) Si estandarizamos la puntuación de 285, resulta un valor *z* de –4,74, lo que supone (mirando las tablas de la normal) aproximadamente que el 0% es la probabilidad de obtener dicha puntuación.

$$p(X < 285) = p(Z < \frac{285 - 300}{10/\sqrt{10}}) = p(Z < -4{,}74) \approx 0$$

b) El intervalo de confianza es:

$$z_{\alpha/2,} = 1{,}64$$

$$I = 285 \pm 1{,}64 \cdot \frac{10}{\sqrt{10}} = 285 \pm 5{,}17 \qquad [279{,}83;\ 290{,}19]$$

c) 300 está fuera del intervalo y, por lo tanto, con un nivel de confianza del 90%, podremos afirmar que la media no llega a 300 ejemplares, sino que está por debajo.

5) La hipótesis nula es $\sigma^2 = 1{,}7$ y la alternativa es $\sigma^2 < 1{,}7$.

El estadístico de contraste es: $\chi^2 = \frac{(400-1)s^2}{1{,}7}$, donde s^2 es la varianza muestral. Entonces

$\chi^2 = 469{,}412$ y su distribución es la de χ^2 la con $400-1 = 399$ grados de libertad.

En este caso, el *p*-valor vale $P(\chi^2 < 469{,}412) = 0{,}991406$ y por lo tanto, no rechazamos la hipótesis nula: no podemos afirmar que sea inferior a 1,7. El valor crítico es 363,253.

6)

a) Hemos de hacer el contraste de una media con varianza desconocida. Las hipótesis nula y alternativa son: $\begin{matrix} H_0 : \mu = 30 \\ H_1 : \mu > 30 \end{matrix}$, donde μ representa la media del tiempo de transferencia de un fichero de 2Mb. El estadístico de contraste es $t = 0{,}378$. El valor crítico valdrá: $t_{0.05,11} = 1{,}80$. Como que $t < t_{0.05,11}$, aceptamos la hipótesis nula y concluimos que la afirmación del responsable es cierta. Si quisiéramos hallar el *p*-valor, éste sería: $p = p(t_{11} > 0{,}378) \approx 0{,}36$. Como es un *p*-valor alto, mayor que 0,05, aceptamos la hipótesis nula tal y como hemos hecho antes.

b) Hemos de hacer el contraste de una media con varianza conocida.

La hipótesis nula y la alternativa son: $\begin{matrix} H_0 : \mu = 30 \\ H_1 : \mu > 30 \end{matrix}$, donde μ representa la media del tiempo de transferencia de un fichero de 2Mb.

El estadístico de contraste es: $z = \frac{\overline{x} - 30}{\sigma/\sqrt{12}}$, donde $\overline{x}$ es la media muestral y σ es la desviación estándar poblacional. La distribución de *z* es la de una normal *N*(0,1). La media y la desviación estándar poblacionales valen respectivamente: $\overline{x} = 30{,}2,\ \sigma = \sqrt{9{,}2} \approx 3{,}03$. El valor del estadístico de contraste es: $z \approx 0{,}228$.

El valor crítico valdrá: $z_{0,05} \approx 1{,}645$. Como $z < z_{0,05}$, volvemos a aceptar la hipótesis nula y concluimos que la afirmación del responsable no es cierta. Si quisiéramos hallar el *p*-valor, éste sería: $p = p(z > 0{,}228) \approx 0{,}41$. Como es un *p*-valor alto, mayor que 0,05, aceptamos la hipótesis nula como hemos hecho anteriormente. Por tanto, hemos llegado a la misma conclusión que en el apartado anterior.

Inferencia de información para dos o más poblaciones

Contrastes de hipótesis para dos poblaciones y comparación de grupos mediante ANOVA

Blanca de la Fuente y Ángel A. Juan

PID_00161060

UOC
Universitat Oberta de Catalunya
www.uoc.edu

Índice

Introducción

En los módulos anteriores se introdujeron los conceptos básicos de estimación y de contraste de hipótesis relacionados con una población. En la práctica cotidiana, sin embargo, es fácil encontrarse con situaciones en las que se dispone de dos o más grupos de individuos o poblaciones y, en tal caso, el interés radica a menudo en ser capaz de discernir si dichos grupos o poblaciones se pueden considerar como semejantes –desde un punto de vista estadístico– o si, por el contrario, son grupos o poblaciones que muestran diferencias significativas entre ellos. Así, por ejemplo, puede ser conveniente comparar las calificaciones medias de dos grupos de estudiantes en función de si han hecho o no uso de una metodología docente innovadora, comparar los porcentajes de recuperación de dos o más grupos de enfermos según el tratamiento recibido, comparar las calidades medias de diferentes accesos a Internet en función de la empresa proveedora, comparar los precios medios de los servicios de obtención de documentos en función de la institución que los ofrezca, etc.

Cuando se consideran dos grupos o poblaciones, las técnicas que se usan para comparar las respectivas medias o proporciones son muy similares a las utilizadas en el caso de una población: contrastes de hipótesis basados en el uso de la distribución normal (cuando se comparan dos proporciones) o de la *t*-Student (cuando se comparan dos medias). En el caso de la comparación entre dos medias de grupos distintos, hay que distinguir si se trata de dos grupos independientes (por ejemplo, cuando se comparan los resultados de un test realizados a dos grupos distintos de individuos) o bien si se trata de dos grupos dependientes (por ejemplo, cuando se están considerando los resultados de un test previo con los resultados de un test posterior, ambos realizados al mismo grupo de individuos).

Finalmente, en el caso de que se deseen comparar más de dos grupos o poblaciones, los contrastes anteriores ya no sirven y resulta necesario recurrir a las técnicas ANOVA basadas en la distribución *F*-Snedecor. El uso de estas técnicas posibilita discernir si las medias correspondientes a un conjunto de tres o más grupos son todas aproximadamente iguales o si, por el contrario, se puede establecer que existen diferencias significativas entre algunas de ellas (y, por consiguiente, entre los grupos asociados).

Objetivos

Los objetivos académicos del presente módulo se describen a continuación:

1. Comparar dos poblaciones utilizando procedimientos similares a los vistos para una sola población.

2. Aprender a formular una hipótesis sobre la naturaleza de las dos poblaciones y la diferencia entre sus medias o proporciones.

3. Conocer el método para comparar las varianzas de dos poblaciones. Para realizar estos contrastes se introduce la distribución F.

4. Entender la importancia práctica de las técnicas ANOVA a la hora de discernir si existen diferencias significativas entre más de dos grupos o poblaciones.

5. Aprender a usar los tests F de ANOVA y saber interpretar adecuadamente los resultados que ofrecen.

6. Comprender la lógica que subyace a la metodología ANOVA.

7. Conocer las hipótesis que se han de satisfacer para poder aplicar las técnicas ANOVA con garantías.

8. Aprender a usar software estadístico y/o de análisis de datos como instrumento básico en la aplicación práctica de los conceptos y técnicas estadísticas.

1. Contrastes de hipótesis para dos poblaciones

En este módulo se presentan métodos para contrastar las diferencias entre las medias o proporciones de dos poblaciones y para contrastar varianzas.

Para comparar las medias o las proporciones poblacionales, se extrae una muestra aleatoria de las dos poblaciones y la inferencia sobre la diferencia entre ambas medias o proporciones se basa en los resultados muestrales. El método apropiado para analizar la información depende del procedimiento empleado al seleccionar las muestras. Consideramos las dos posibilidades siguientes:

a) Muestras dependientes (**datos pareados**): en este procedimiento, las muestras se eligen por pares, una de cada población. La idea es que aparte de la característica objeto del estudio, los elementos de cada uno de estos pares deben estar relacionados, de manera que la comparación pueda establecerse directamente. Por ejemplo, supongamos que queremos medir la eficacia de un curso de lectura rápida. Una manera de abordar el problema sería tomar nota de las palabras leídas por minuto por una muestra de alumnos antes de tomar el curso y compararlas con los resultados obtenidos *por los mismos alumnos* una vez completado el curso. En este caso, cada par consistiría en medidas de velocidad de un mismo alumno realizadas antes y después del curso, se podría averiguar si existen pruebas contundentes de la eficacia del curso de lectura rápida.

b) Muestras independientes: en este método se extraen muestras independientes de cada una de las dos poblaciones, de manera que los miembros de una muestra no tienen necesariamente relación con los miembros de la otra. Por ejemplo, se realiza un estudio para evaluar las diferencias en los niveles educativos entre dos centros de capacitación, se aplica un examen común a personas que asisten a cada centro. Las calificaciones del examen son uno de los factores principales para evaluar diferencias de calidad entre los centros.

1.1. Contrastes de hipótesis para la diferencia de medias

Contraste de hipótesis para la diferencia entre las medias de dos poblaciones: muestras independientes.

En este apartado presentaremos los procedimientos para contrastar las hipótesis acerca de la diferencia de medias de dos poblaciones.

Se supone que se dispone de muestras aleatorias independientes de n_1 y n_2, observaciones procedentes de dos poblaciones normales con medias μ_1 y μ_2 y varianzas conocidas σ_1^2 y σ_2^2 respectivamente. Se desea contrastar la hipótesis nula (H_0) que afirma que los valores de las medias de las dos poblaciones son iguales: $\boldsymbol{H_0: \mu_1 - \mu_2 = 0}$ frente a cualquiera de las hipótesis alternativas:

Nota

A veces en lugar de:
H_0: $\mu_1 - \mu_2 = 0$
escribiremos:
H_0: $\mu_1 = \mu_2$

H_1: $\mu_1 - \mu_2 \neq 0$, H_1: $\mu_1 - \mu_2 < 0$, H_1: $\mu_1 - \mu_2 > 0$. Se fija un nivel de significación α para realizar el contraste.

El estadístico de contraste será:

$$z^* = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}},$$

donde $\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} = \sigma_{\bar{x}_1 - \bar{x}_2}$ es el error estándar.

Es una observación de una distribución $N(0,1)$.

Recordad

$\bar{X}_1 \rightarrow N(\mu_1, \sigma_1)$ y
$\bar{X}_2 \rightarrow N(\mu_2, \sigma_2)$

La variable diferencia de medias muestrales:

$$(\bar{X}_1 - \bar{X}_2) \rightarrow N\left(\mu_1 - \mu_2, \sqrt{\sigma_1^2 + \sigma_2^2}\right)$$

En el caso de que no se pueda asegurar que las muestras provienen de poblaciones normales, sólo podremos contrastar la diferencia de medias si los tamaños de las muestras son superiores a treinta.

El teorema central del límite dice que si tenemos un grupo numeroso de variables independientes y todas ellas siguen el mismo modelo de distribución (cualquiera que éste sea), la suma de ellas se distribuye según una distribución normal estándar.

Por lo tanto el estadístico de contraste:

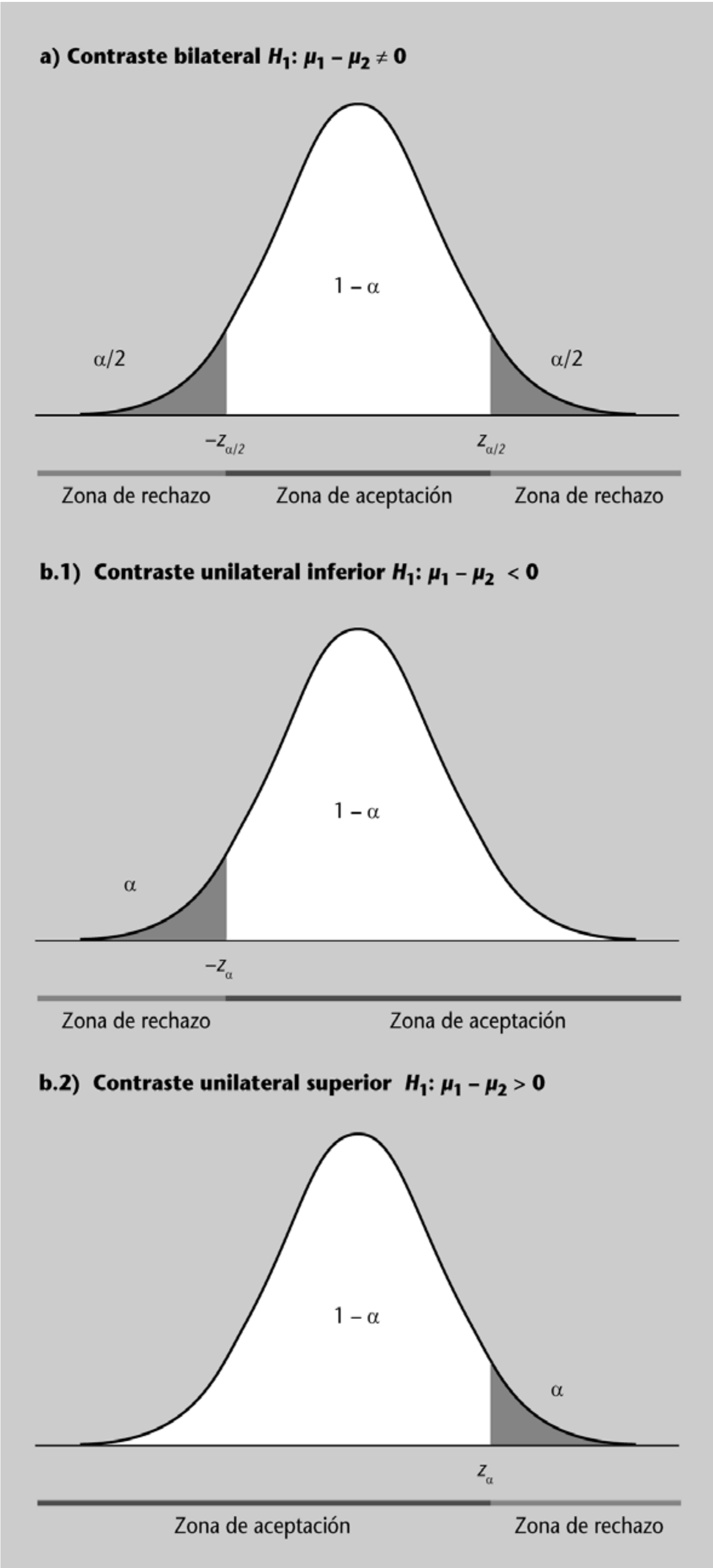
$$z^* = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

es una observación de una variable aleatoria que se distribuye aproximadamente como una $N(0,1)$.

Regla de decisión del contraste de hipótesis

Las regiones de rechazo de la hipótesis nula H_0: $\mu_1 - \mu_2 = 0$ son:

Figura 1. Regiones de rechazo para contrastes de las diferencias de medias



Varianzas poblacionales conocidas

Se puede actuar de dos maneras:

1) A partir del ***p*-valor** según sea H_1:

- p-valor = $P(|Z| > |z^*|)$
- p-valor = $P(Z < z^*)$
- p-valor = $P(Z > z^*)$

2) Si p-valor $\leq \alpha$ **se rechaza H_0** a partir de los valores críticos según sea H_1:

- Si $|z^*| > z_{\alpha/2}$ se rechaza H_0
- Si $z^* < -z_{\alpha}$ se rechaza H_0
- Si $z^* > -z_{\alpha}$ se rechaza H_0

donde

z_{α} es tal que $P(Z > z_{\alpha}) = \alpha$ y

$z_{\alpha/2}$ es tal que $P(Z > z_{\alpha/2}) = \alpha/2$

Una vez que se ha calculado el valor del estadístico de contraste, se debe determinar el *p*-valor. El *p*-valor depende de la hipótesis alternativa planteada.

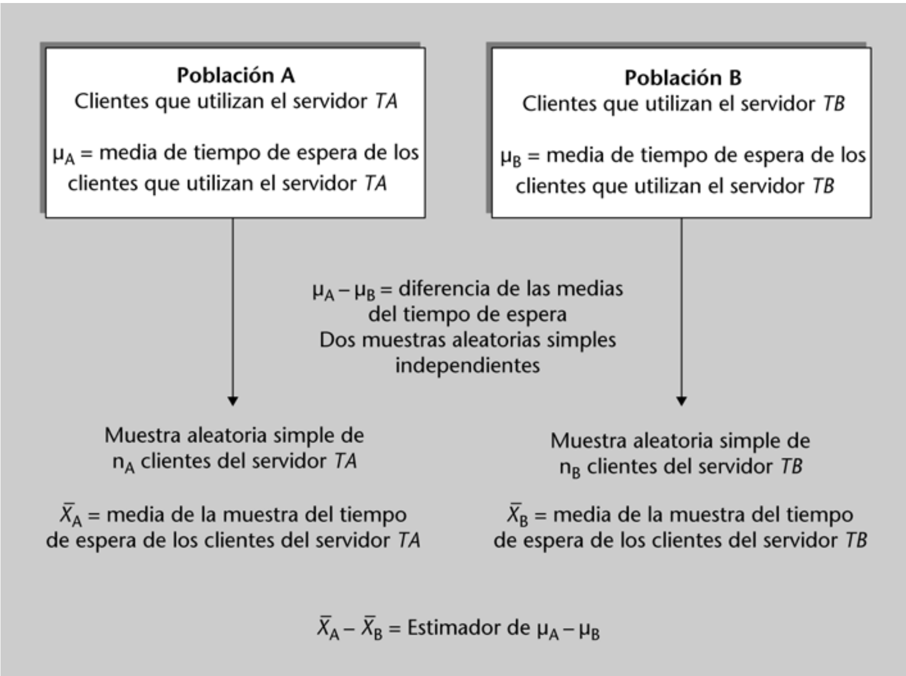
- Si $H_1 : \mu_1 - \mu_2 \neq 0$, entonces $p = 2P(Z < |z|)$
- Si $H_1 : \mu_1 - \mu_2 < 0$, entonces $p = P(Z < z)$
- Si $H_1 : \mu_1 - \mu_2 > 0$, entonces $p = P(Z > z)$

Los *p*-valores de estos contrastes son la probabilidad de obtener un valor al menos tan extremo como el estadístico de contraste obtenido.

Si el *p*-valor es significativo se rechaza la hipótesis nula si es menor que el nivel de significación α fijado.

Ejemplo 1. "Comparación de las medias del tiempo de respuesta de dos servidores".

Figura 2. Estimación de la diferencia entre las medias de dos poblaciones



En una empresa informática se desea medir la eficiencia de dos servidores web. Para ello, miden el tiempo de espera del cliente entre la petición que hace y la respuesta que le da el servidor. En la tabla 1 vemos los tiempos de espera (en milisegundos) de ambos servidores (*TA* y *TB*) para cincuenta peticiones son:

Tabla 1. Datos del ejemplo 1. "Comparación de las medias del tiempo de respuesta de dos servidores"

Tiempo de espera para el servidor *A*				**Tiempo de espera para el servidor *B***			
9,67	10,01	8,08	10,01	6,45	6,94	12,11	10,31
9,62	10,55	9,98	9,96	9,64	10,47	12,55	10,83
9,50	11,26	10,30	9,28	8,53	8,47	7,98	8,41

Tiempo de espera para el servidor *A*				**Tiempo de espera para el servidor *B***			
10,88	10,64	7,05	10,30	9,20	7,42	10,20	9,15
8,94	10,23	11,79	11,08	4,55	7,48	11,28	7,06
10,59	11,63	9,59	10,05	8,51	11,01	6,53	8,04
9,81	8,91	10,88	9,74	12,11	9,56	8,14	11,70
9,46	10,27	9,83	11,14	7,65	6,80	8,99	10,56
9,26	9,49	10,92	9,44	8,85	8,99	10,01	7,82
9,02	8,99	10,98	9,17	8,45	7,48	8,14	6,01
8,61	10,09	9,54	10,86	8,80	12,57	9,69	8,82
9,42	9,11	10,17		8,82	7,97	7,03	
10,86	9,47	10,32		9,85	8,62	8,59	

Supongamos que las muestras aleatorias de los tiempos de espera son independientes. La empresa quiere saber si el servidor *A* es menos eficiente (más lento) que el servidor *B* con un nivel de confianza del 99%.

Para contestar a estas preguntas se hará un contraste para comparar dos medias. Dado que el enunciado nos pregunta "si el servidor *A* es menos eficiente que el servidor *B*", considerando que un servidor es menos eficiente si es más lento, entonces hemos de contrastar si la media del tiempo de espera del servidor *A* es más grande que la media del tiempo de espera del servidor *B*. Así pues, tenemos que plantear una **hipótesis alternativa unilateral**.

- Las hipótesis nula y alternativa son: $H_0 : \mu_A - \mu_B = 0$, $H_1 : \mu_A - \mu_B > 0$
- Fijamos $\alpha = 0{,}01$.
- No podemos asegurar que las poblaciones sean normales, pero como hemos mencionado anteriormente, al tratarse de muestras grandes (superiores a treinta observaciones) el estadístico de contraste será:

$$z^* = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Es una observación de una variable aleatoria que **se distribuye aproximadamente como una *N*(0,1)**.

Para resolverlo manualmente calcularemos primero los valores muestrales como hemos expuesto en los módulos anteriores:

Tiempo de espera para el servidor *A*	Tiempo de espera para el servidor *B*
$n_A = 50$	$n_B = 50$
$\bar{x}_A = 9{,}94$	$\bar{x}_B = 8{,}90$
$s_A = 0{,}90$	$s_B = 1{,}75$

Las varianzas muestrales s_A^2 y s_B^2 para estimar las varianzas poblacionales y calcular el estadístico z^*:

$$z^* = \frac{(9,94 - 8,90)}{\sqrt{\frac{0,90^2}{50} + \frac{1,75^2}{50}}} = 3,75$$

Ahora se puede calcular el p-valor $p = P(Z > 3,75) = 0,00$

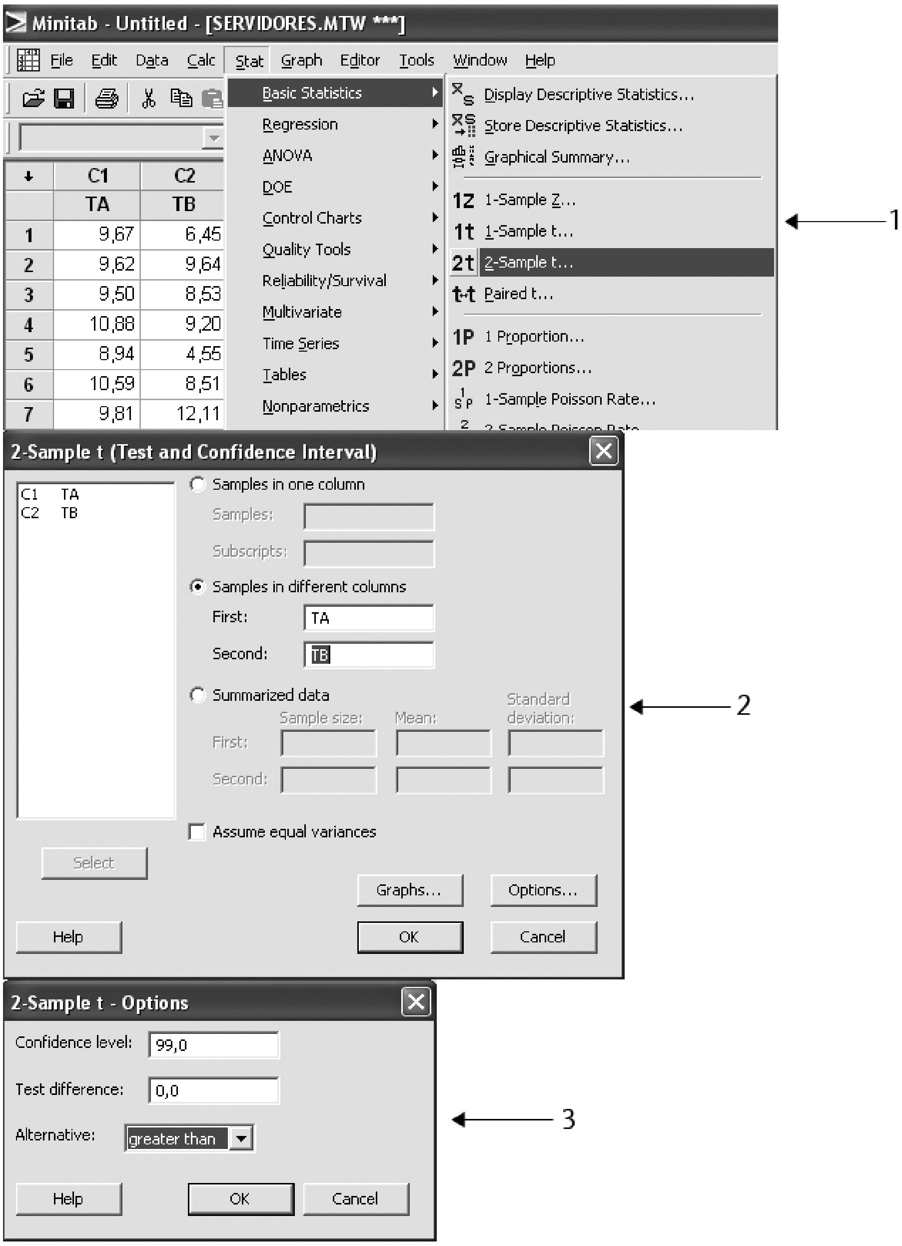
Puesto que el p-valor es menor que $\alpha = 0,01$, se rechaza la hipótesis nula a favor de la alternativa. Así, el tiempo medio de espera del servidor A es mayor que el del B. Luego el servidor A es menos eficiente que el B.

Ejemplo con Minitab: si el ejemplo anterior se resuelve con Minitab, se observa que el programa no ofrece la opción de usar la distribución normal. De todas formas, dado que las muestras son muy grandes, sabemos que la distribución t de Student se acerca a la normal a medida que aumenta el número de grados de libertad. Por tanto, los resultados que da Minitab no serán similares por la aproximación a lo normal.

Los resultados de la figura 3 muestran el p-valor = 0,000 < 0,001. Esto indica que podemos rechazar la hipótesis nula concluyendo que las medias de tiempos de espera del servidor A es mayor que las del B. Luego el servidor A es menos eficiente que el B.

Los grados de libertad (DF) del estadístico t aumentan si las poblaciones tienen distribución aproximadamente normal pero las varianzas poblacionales no son iguales.

Figura 3. Pasos para realizar un contraste de hipótesis para la diferencia de medias para muestras independientes



Pasos a seguir

Una vez introducidos los datos en el programa, se sigue la ruta ***Stat*** > ***Basic Statistics*** > ***2-Sample t (1)***, y se seleccionan las variables en la ventana correspondiente ***(2)***. En el cuadro de dialogo ***Options*** se completan los campos ***Confidence level***: 99,0 y el tipo de hipótesis alternativa ***Alternative:*** *greater than* ***(3)***.

Seleccionad ***OK*** para obtener el contraste.

Observad

En el paso ***(2)*** no presuponemos que las varianzas sean iguales.

Figura 4. Resultados del contraste de hipótesis

Two-Sample T-Test and CI: TA; TB

```
Two-sample T for TA vs TB

     N   Mean  StDev  SE Mean
TA  50  9,935  0,900     0,13
TB  50   8,90   1,75     0,25


Difference = mu (TA) - mu (TB)
Estimate for difference:  1,032
99% lower bound for difference:  0,371
T-Test of difference = 0 (vs >): T-Value = 3,71
 P-Value = 0,000  DF = 73
```

Contrastes para muestras con varianzas poblacionales desconocidas pero iguales

El procedimiento que utilizamos se basa en la distribución t con $n_1 + n_2 - 2$ grados de libertad.

El estadístico de contraste será:

$$t = \frac{(\bar{x}_1 - \bar{x}_2)}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

donde s es la desviación típica común.

Ejemplo 2. "Estudio sobre la producción científica".

El director de una escuela universitaria quiere comparar dos departamentos, *A* y *B*, de tamaño similar, por lo que se refiere al "número total de publicaciones o ponencias de calidad que puedan aportar mejoras a la actividad docente de la escuela". Se considerará que una publicación es de calidad cuando se haya publicado en una revista indexada o la haya publicado una editorial de prestigio internacional; se considerará que una ponencia es de calidad cuando se haya desarrollado en un congreso internacional con proceso de selección; para determinar si la publicación o ponencia puede aportar mejoras a la actividad docente se ha constituido un tribunal de expertos independientes.

Se ha tomado una muestra aleatoria formada por seis profesores del departamento *A* y se ha hallado el valor de la variable "número total de publicaciones o ponencias de calidad para cada uno de dichos profesores". Se ha hecho lo propio con otra muestra aleatoria formada por ocho profesores del departamento *B*. Los resultados se presentan a continuación:

Tabla 2. Datos del ejemplo 2. "Estudio sobre la difusión científica"

Dep. *A*	5	8	7	6	9	7		
Dep. *B*	8	10	7	11	9	12	14	9

Para un nivel de significación alfa = 0,05, ¿puede afirmarse que la producción media de ambos departamentos (según los criterios establecidos) es significativamente distinta?

Para realizar el estudio partiremos del supuesto de que no hay diferencias en el "número total de publicaciones o ponencias de calidad de ambos departamentos". Por consiguiente, en términos de la media del número total de publicaciones o ponencias de calidad, la hipótesis nula es que la diferencia de medias es cero. Si la evidencia de la muestra conduce al rechazo de esta hipótesis, llegaremos a la conclusión de que las medias de calidad son distintas para las dos poblaciones, lo que indica que hay diferencia en las publicaciones de calidad de los dos departamentos, y eso induciría a encontrar las razones de esa diferencia.

En este estudio hay dos poblaciones: una de los profesores del departamento *A*, y otra de los profesores del departamento *B*. Suponemos que ambas poblaciones son normales y que sus varianzas son iguales pero desconocidas.

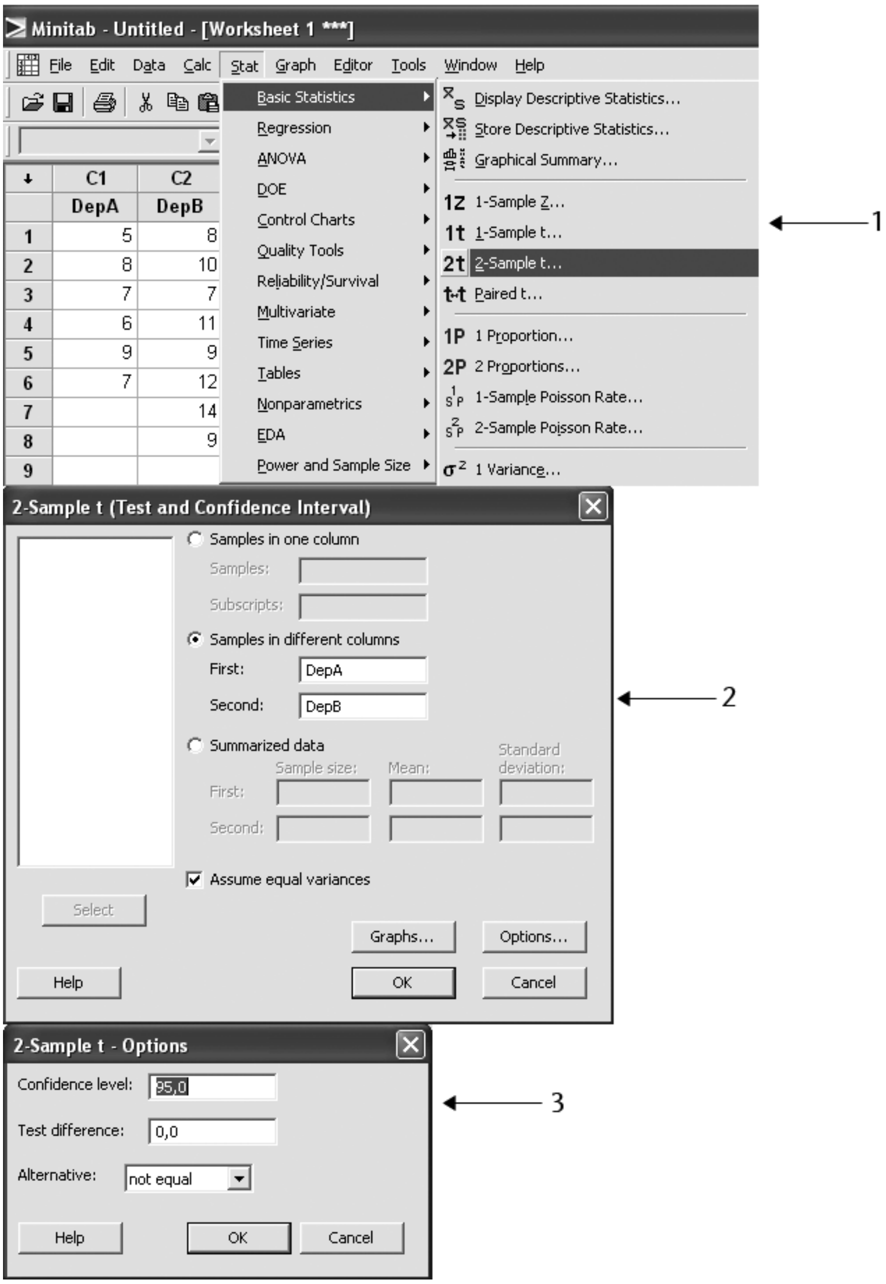
Considerando el número de publicaciones y ponencias, las medias de población son: μ_A y μ_B, se plantean las hipótesis de trabajo:

$$H_0\text{: } \mu_1 - \mu_2 = 0 \text{ (ambas medias son iguales)}$$
$$H_1\text{: } \mu_1 - \mu_2 \neq 0 \text{ (ambas medias son distintas)}$$

Se trata de un contraste de **hipótesis bilateral** sobre la media de dos poblaciones independientes.

Se empleara Minitab para probar las hipótesis acerca de la diferencia entre las medias de dos poblaciones (figura 5).

Figura 5. Pasos para realizar un contraste de hipótesis para la diferencia de medias para muestras con varianzas poblacionales desconocidas



Pasos a seguir

Una vez introducidos los datos en el programa se sigue la ruta ***Stat*** > ***Basic Statistics*** > ***2-Sample t (1)***, y se seleccionan las variables en la ventana correspondiente ***(2)***.En el cuadro de dialogo ***Options*** se completan los campos ***Confidence level***: 95,0 y el tipo de hipótesis alternativa ***Alternative***: *not equal* ***(3)***.

Seleccionad ***OK*** para obtener el contraste.

Observad

En el paso ***(2)*** suponemos que las varianzas son desconocidas pero iguales y marcamos la casilla correspondiente.

Obtuvimos los resultados de la figura 6. Aparecen los valores muestrales de ambos departamentos. El estadístico de contraste es un valor $t = -2{,}84$ con 12 grados de libertad (*DF*) y el *p*-valor *P-Value* = 0,015.

Figura 6. Resultados del contraste de hipótesis

```
Two-Sample T-Test and CI: DepA; DepB

Two-sample T for DepA vs DepB

      N   Mean  StDev  SE Mean
DepA  6   7,00   1,41     0,58
DepB  8  10,00   2,27     0,80


Difference = mu (DepA) - mu (DepB)
Estimate for difference:  -3,00
95% CI for difference:  (-5,30; -0,70)
T-Test of difference = 0 (vs not =): T-Value = -2,84
P-Value= 0,015  DF = 12
Both use Pooled StDev = 1,9579
```

Como *p*-valor = 0,015 < 0,05, se puede rechazar la hipótesis nula con $\alpha = 0{,}05$. Así, la producción media de ambos departamentos (según los criterios establecidos) es significativamente distinta en los departamento *A* y *B*. Observad que la información de Minitab para el intervalo de confianza del 95% en la figura 5 tiene como extremos los valores –5,30 y –0,70 (observad que el 0 no está incluido en dicho intervalo). Esto también nos indica que debemos rechazar la hipótesis nula y aceptar la alternativa (las medias son distintas).

Así, los resultados permiten que el director de la escuela universitaria concluya que existen diferencias significativas entre ambos departamentos en el "número total de publicaciones o ponencias de calidad".

Aplicando **Microsoft Excel** al ejemplo 2. "Estudio sobre la difusión científica".

Para ejecutar una prueba *t* de dos muestras independientes para datos no apareados haced clic en (*t-Test: Two Simple* > *Assuming Equal Variants*) "prueba *t*: dos muestras suponiendo varianzas iguales" y especificad las dos columnas que contengan los datos.

La figura 7 muestra el correspondiente *output* que ofrece **Microsoft Excel**.

Figura 7. Resultados ejemplo 2. "Estudio sobre la difusión científica". Excel

	A	B	C
1	Prueba *t* para dos muestras suponiendo varianzas iguales		
2			
3		*DepA*	*DepB*
4	Media	7	10
5	Varianza	2	5,14285714
6	Observaciones	6	8
7	Varianza agrupada	3,83333333	
8	Diferencia hipotética de las medias	0	
9	Grados de libertad	12	
10	Estadístico *t*	-2,8371975	
11	*P*(*T*<=*t*) una cola	0,0074872	
12	Valor crítico de *t* (una cola)	1,78228755	
13	*P*(*T*<=*t*) dos colas	0,01497439	
14	Valor crítico de *t* (dos colas)	2,17881283	

Análisis de datos

Para realizar contrastes de hipótesis con **MS Excel** es necesario instalar previamente un complemento llamado "Análisis de datos". Para instalar las herramientas de análisis de datos haced clic en ***Herramientas*** > ***complementos***, en el cuadro de diálogo activar ***Herramientas para análisis***.

Como observamos, el p-valor = 0,0149, al ser menor que el valor de α, se puede rechazar la hipótesis nula con α = 0,05.

Contraste de hipótesis para la diferencia entre las medias de dos poblaciones: muestras dependientes (datos pareados)

Disponemos de una muestra aleatoria de n pares de observaciones de distribuciones con medias μ_A y $\mu_{B.}$ Denotamos por $\bar{d}$ y s_d la media muestral y la desviación típica observadas para las n diferencias (x_A-x_B) y sea $\mu_d = \mu_A - \mu_B$ media de las diferencias para la población.

Si la distribución poblacional es normal podemos realizar los siguientes contrastes para un nivel de significación α:

la hipótesis nula: $H_0 = \mu_d = 0$

la hipótesis alternativa (H_1) puede ser bilateral: $H_1 : \mu_d \neq 0$

o unilateral H_1: $\mu_d > 0$ o $H_1 : \mu_d < 0$

En este tipo de contraste se usa la misma metodología usada para el contraste de la media para una sola población que vimos en el módulo anterior.

Para ilustrar el diseño con muestras emparejadas ilustrar el ejemplo siguiente:

Ejemplo 3. "Puntuaciones de un test de actitud"

A un grupo de personas se les propuso un test de actitud acerca de un tema polémico y obtuvimos unos resultados. Luego el grupo asistió a la proyección de una película favorable al tema y acto seguido se les propuso de nuevo el test de actitud, del que se obtuvieron otros resultados. En la tabla 3 aparecen los datos acerca de las puntuaciones del test realizado a once personas. Cada persona da un par de valores, uno para antes de asistir a la proyección de la película y otro después de asistir a la proyección. Se quiere verificar la hipótesis de que la proyección de una película favorable hace que cambie la actitud desfavorable hacia el tema.

Tabla 3. Datos del ejemplo 3. "Puntuaciones de un test de actitud"

Persona	Puntuación del test antes de ver la película	Puntuación del test después de ver la película	Diferencia de puntuaciones del test (d_i)
1	24	16	8
2	20	18	2
3	24	20	4
4	28	24	4
5	30	24	6

Muestras dependientes

Muestras dependientes significa que tenemos **una muestra** de observaciones de dos variables.

La media de la muestra es:

$$\bar{d} = \frac{\sum_i^n d_i}{n}$$

La desviación estándar es:

$$s_d = \sqrt{\frac{\sum_i^n (d_i - \bar{d})^2}{n-1}}$$

La notación d es para recordar que la muestra apareada produce datos de *diferencia.*

Persona	Puntuación del test antes de ver la película	Puntuación del test después de ver la película	Diferencia de puntuaciones del test (d_i)
6	20	22	-2
7	24	20	4
8	22	18	4
9	18	10	8
10	18	8	10
11	24	20	4
			$\sum_{i=1}^{11} d_i = 52$

Observad que la última columna de la tabla 3 contiene la diferencia entre las puntuaciones antes y después de ver la película. La clave para analizar el diseño con muestras apareadas es tener en cuenta que sólo se considera la columna de las diferencias. Verificaremos la hipótesis de investigación a un nivel de significación del 1% ($\alpha = 0{,}01$). Sea μ_d = la media de las diferencias para la población de personas.

Las hipótesis serán:

$$H_0\text{: } \mu_d = 0$$
$$H_1\text{: } \mu_d \neq 0$$

Se trata de un contraste bilateral. Si se rechaza H_0 se llega a la conclusión de que las medias de las puntuaciones del test son distintas al nivel de significación del 1%. En el módulo 2 se vio que si se puede suponer que la población tiene una distribución normal, el estadístico de contraste es una ***t*-Student** con $n - 1$ grados de libertad, para probar la hipótesis nula acerca de la media poblacional, si no conocemos la varianza de la población como en este ejemplo.

Con datos de diferencia se calcula el estadístico de prueba para la hipótesis nula

$$H_0\text{: } \mu_d = 0 \text{ es:}$$

$$\text{como } \bar{d} = \frac{52}{11} = 4{,}72 \text{ y } s_d = \sqrt{\frac{106{,}18}{10}} = 3{,}26$$

$$t^* = \frac{\bar{d} - \mu_d}{s_d / \sqrt{n}} = \frac{4{,}72 - 0}{3{,}26 / \sqrt{11}} = 4{,}80$$

Con $\alpha = 0{,}01$ y $n - 1 = 10$ grados de libertad ($t_{0,01/2} = t_{0,005} = 3{,}169$), la regla de rechazo para la prueba bilateral es:

$$\text{Rechazar } H_0 \text{ si } t^* < 3{,}169 \text{ o } t^* > 3{,}169$$

En vista de que $t^{*} = 4{,}80$ está en la región de rechazo, se rechaza H_0 y se acepta H_1 y podemos afirmar con un 99% de confianza que la película influyó en la actitud de las personas.

Con los resultados de la muestra podemos definir un intervalo de confianza de diferencia entre las dos medias de la población, con la metodología para población única del módulo 2 los cálculos son:

$$\overline{d} \pm t_{\alpha/2} \frac{s_d}{\sqrt{n}} = 4{,}72 \pm 3{,}169\left(\frac{3{,}26}{\sqrt{11}}\right) = 4{,}72 \pm 3{,}12 = [1{,}60;\ 7{,}84]$$

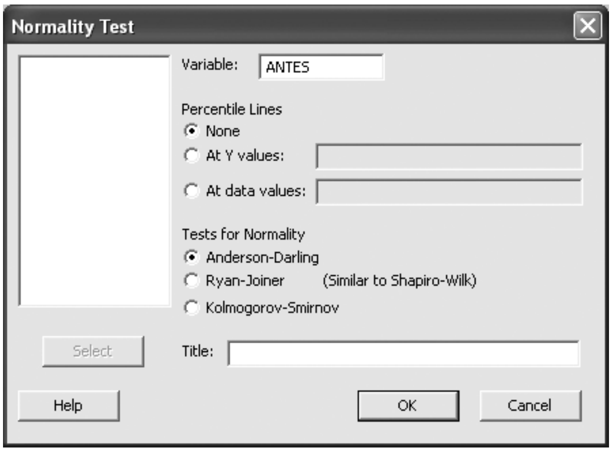
En consecuencia, el intervalo de confianza de 99% de la diferencia de medias entre las medias de las dos puntuaciones del test es de 1,6 hasta 7,84 observamos que el intervalo no incluye el valor cero, luego, como hemos visto en el contraste, podemos rechazar H_0.

Emplearemos Minitab para este ejemplo 3. "Puntuaciones de un test de actitud".

La figura 8 muestra los pasos básicos necesarios para realizar el contraste de hipótesis.

En primer lugar comprobaremos el supuesto de que las poblaciones siguen una distribución aproximadamente normal:

Figura 8. Pasos para realizar un test de normalidad. Minitab



Pasos a seguir

Una vez introducidos los datos en el programa se sigue la ruta ***Stat*** > ***Basic Statistics*** > ***Normality Test***. Y rellenamos los campos en la ventana correspondiente.

En el cuadro de diálogo se selecciona el test de ***Anderson-Darling***.

En los gráficos resultantes (figuras 9 y 10) se observa que no hay indicios para dudar de que se cumpla el supuesto de normalidad, ya que los puntos se encuentran muy próximos a las respectivas rectas. Los gráficos nos proporcionan también el p-valor asociado al **test de normalidad de Anderson-Darling**, siendo dicho p-valor suficientemente grande en ambos casos para no descartar la hipótesis nula de este contraste: que los datos siguen una distribución normal.

Figura 9. Test de normalidad. Minitab

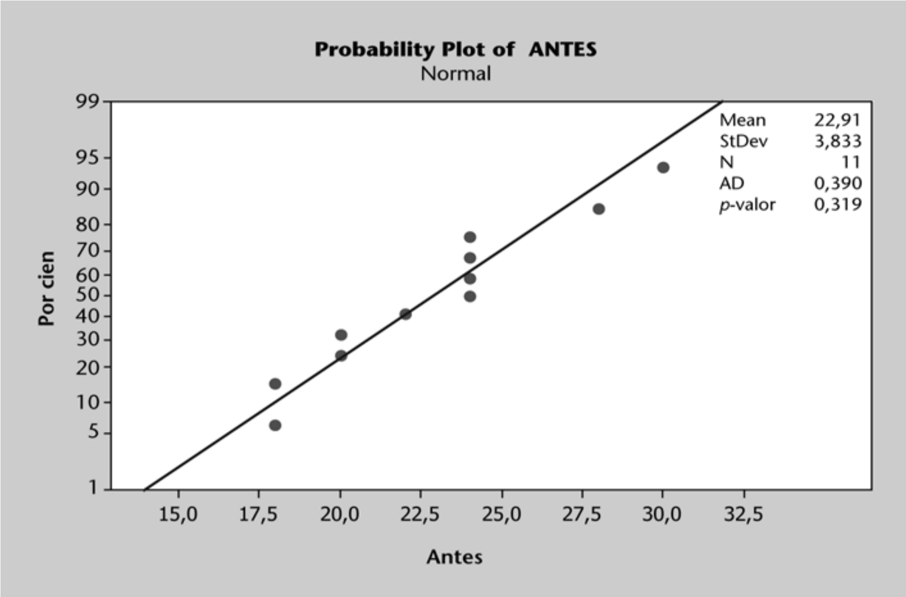
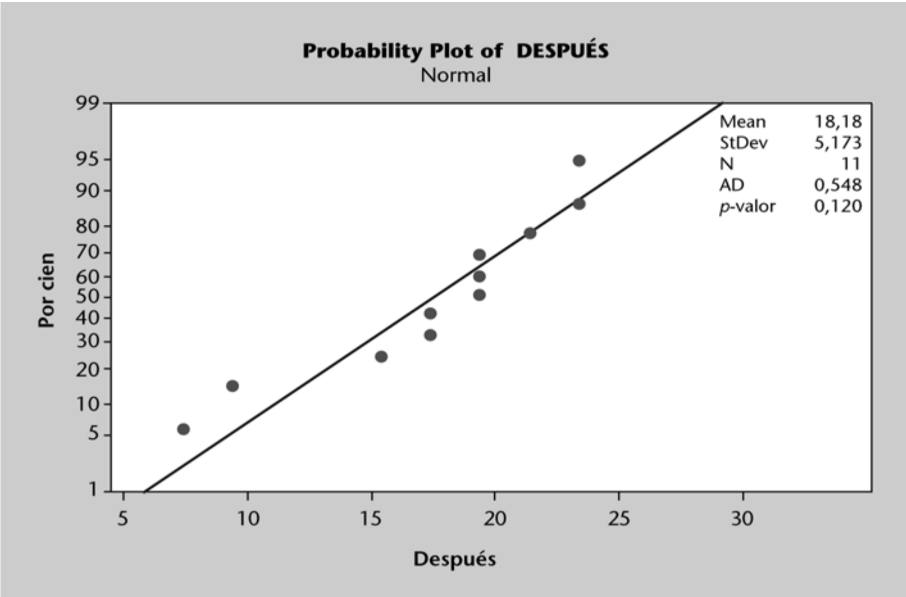
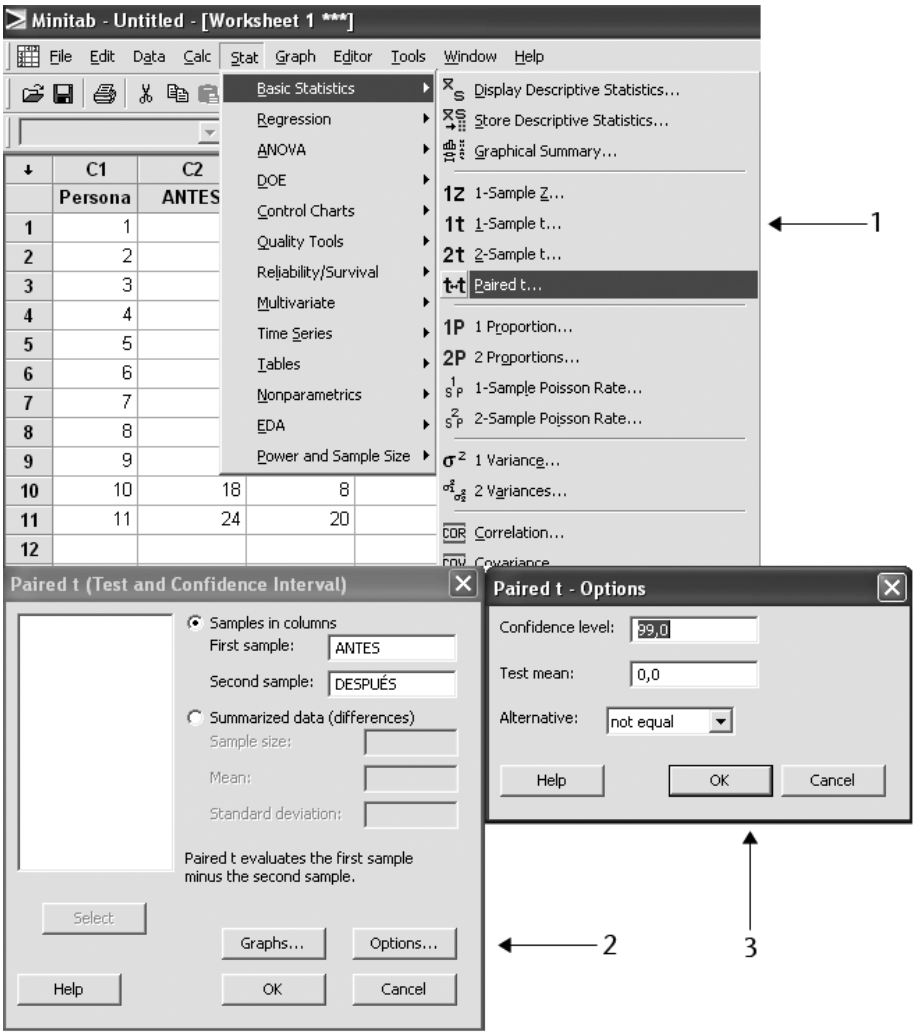


Figura 10. Test de normalidad. Minitab



Pasamos, pues a realizar las inferencias ya comentadas sobre μ_d.

Figura 11. Pasos para realizar un contraste de hipótesis para la diferencia de medias para muestras dependientes



Pasos a seguir

Se sigue la ruta ***Stat*** > ***Basic Statistics*** > ***Paired t (1)*** y se rellenan los campos en la ventana correspondiente ***(2)***. En el cuadro de diálogo ***Options*** se completan los campos ***Confidence level***: 99,0 y el tipo de hipótesis alternativa ***Alternative***: *not equal* ***(3)***.

Seleccionad ***OK*** para obtener el contraste.

Los resultados obtenidos en la figura 12 que, en base a las observaciones registradas, hay una probabilidad de 0,99 de que μ_d sea un valor del intervalo (1,613; 7,841). Además, con un *p*-valor de 0,001 también podemos afirmar que hay indicios suficientes para rechazar la hipótesis nula. Por lo tanto, podemos concluir que la película influyó en la actitud de las personas.

Figura 12. Resultados del contraste de medias para dos muestras dependientes. Minitab

```
Paired T for ANTES - DESPUÉS

             N   Mean  StDev  SE Mean
ANTES       11  22,91   3,83     1,16
DESPUÉS     11  18,18   5,17     1,56
Difference  11  4,727  3,259    0,982


99% CI for mean difference: (1,613; 7,841)
T-Test of mean difference = 0 (vs not = 0): T-Value = 4,81
P-Value = 0,001
```

También puede ejecutar una prueba *t* por pares utilizando **Excel**. Desde *Herramientas* > *Análisis de datos*, haced clic en *Prueba t para medias de dos muestras emparejadas* y especificad las dos columnas que contienen los datos por pares. Este comando no calcula el intervalo de confianza, de modo que tenéis que calcularlo mediante las fórmulas que aparecen en este módulo.

La figura 13 muestra el correspondiente *output* que ofrece **Microsoft Excel**.

Figura 13. Resultados del contraste de medias para dos muestras emparejadas. Excel

B13 fx 0,000711223615253786

	A	B	C
1	Prueba *t* para medias de dos muestras emparejadas		
2			
3		*ANTES*	*DESPUÉS*
4	Media	22,90909091	18,18181818
5	Varianza	14,69090909	26,76363636
6	Observaciones	11	11
7	Coeficiente de correlación de Pea	0,777564218	
8	Diferencia hipotética de las media	0	
9	Grados de libertad	10	
10	Estadístico *t*	4,811515866	
11	*P*(*T*<=*t*) una cola	0,000355612	
12	Valor crítico de *t* (una cola)	2,763769458	
13	*P*(*T*<=*t*) dos colas	0,000711224	
14	Valor crítico de *t* (dos colas)	3,169272672	

Al ser el p-valor = 0,0007 < α(0,01), se rechaza H_0.

1.2. Contrastes de hipótesis para la diferencia de proporciones

Al estudiar la diferencia entre dos proporciones poblacionales, el estimador es $\hat{p}_1 - \hat{p}_2$. Como hemos visto en casos anteriores, la distribución del estimador de las muestras es un factor clave para determinar los intervalos de confianza y probar las hipótesis de los parámetros.

Supongamos que disponemos de dos muestras aleatorias simples e independientes de n_1 y n_2 observaciones. Las proporciones muestrales de éxitos son respectivamente: $\hat{p}_1$ y $\hat{p}_2$.

La distribución de la variable diferencia de proporciones muestrales $\hat{p}_1 - \hat{p}_2$ se puede aproximar con una distribución $N(0,1)$.

Bajo el supuesto de la hipótesis nula cierta (H_0: $p_1 - p_2 = 0$), tenemos que el estadístico de contraste es:

$$z^* = \frac{(\hat{p}_1 - \hat{p}_2)}{\sqrt{\dfrac{\hat{p}(1-\hat{p})}{n_1} + \dfrac{\hat{p}\ (1-\hat{p})}{n_2}}}$$

donde ($\hat{p}_1 - \hat{p}_2$) es la diferencia de las proporciones muestrales.

El valor $\hat{p}$ es el valor estimado común de la proporción poblacional, que podemos estimarlo a partir de las dos muestras:

Recordad

Si los tamaños de las muestras son grandes:

$$\hat{p}_1 \rightarrow N\left(p_1, \sqrt{\frac{p_1(1-p_1)}{n_1}}\right)$$

y

$$\hat{p}_2 \rightarrow N\left(p_2, \sqrt{\frac{p_2(1-p_2)}{n_2}}\right)$$

$$\hat{p} = \frac{n_1 \hat{p}_1 + n_2 \hat{p}_2}{n_1 + n_2}$$

Regla de decisión del contraste de hipótesis

Una vez que se ha calculado el valor del estadístico de contraste, se debe determinar el *p*-valor. El *p*-valor depende de la hipótesis alternativa planteada.

- Si $H_1 : p_1 - p_2 \neq 0$, entonces $p = 2P\ (Z < |z|)$
- Si $H_1 : p_1 - p_2 < 0$, entonces $p = P\ (Z < z)$
- Si $H_1 : p_1 - p_2 > 0$, entonces $p = P(Z > z)$

Si el *p*-valor es significativo se rechaza la hipótesis nula si es menor que el nivel de significación α fijado.

Nota

A veces, en lugar de:
H_0: $p_1 - p_2 = 0$
escribiremos:
H_0: $p_1 = p_2$

Se utilizará el ejemplo del apartado 1.1, tabla 1. Datos del ejemplo 1. “Comparación de las medias del tiempo de respuesta de dos servidores”.

En una empresa informática se desea medir la eficiencia de dos servidores web. Para ello, miden el tiempo de espera del cliente entre la petición que éste hace y la respuesta que le da el servidor. Los tiempos de espera (en milisegundos) de ambos servidores (*TA* y *TB*) para cincuenta peticiones están en la tabla 2.

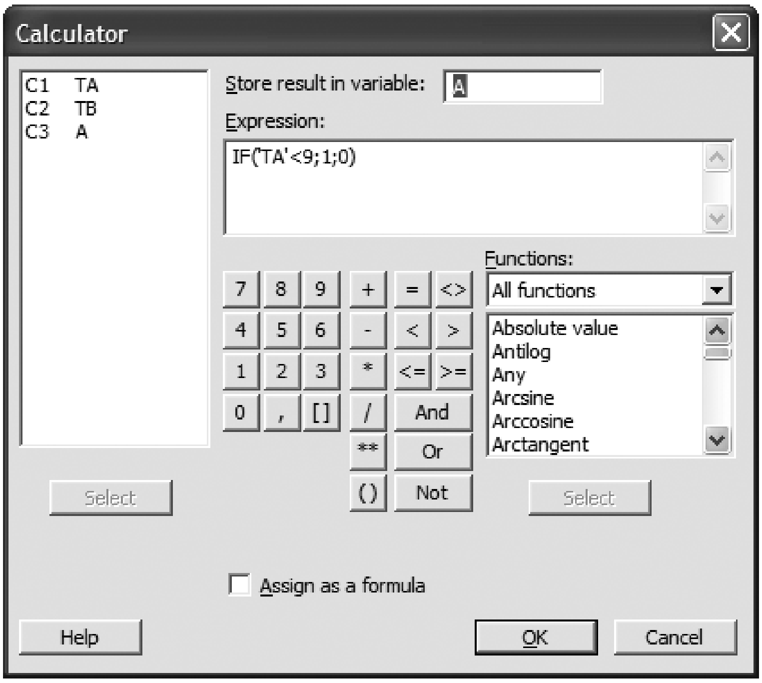
Diremos que el tiempo de espera es aceptable si es menor que 9 milisegundos. ¿Podemos decir que la proporción de peticiones con tiempo de espera aceptable es distinta para los dos servidores?

Para contestar esta pregunta debemos hacer un contraste de diferencia de proporciones que resolveremos con Minitab.

Lo primera operación es calcular para cada tipo de servidor la proporción de tiempo inferior a 9 milisegundos. Para ello, creamos una nueva columna de nombre A donde pondremos un 1 si la observación de tiempo de espera del servidor *A* es inferior a 9 y 0 en caso contrario. Después sumaremos los valores de la columna y obtendremos el número de observaciones de tiempo del servidor *A* inferior a 9 milisegundos.

En la figura 14 se indican los pasos a seguir:

Figura 14. Pasos a seguir para recalcular una variable nueva

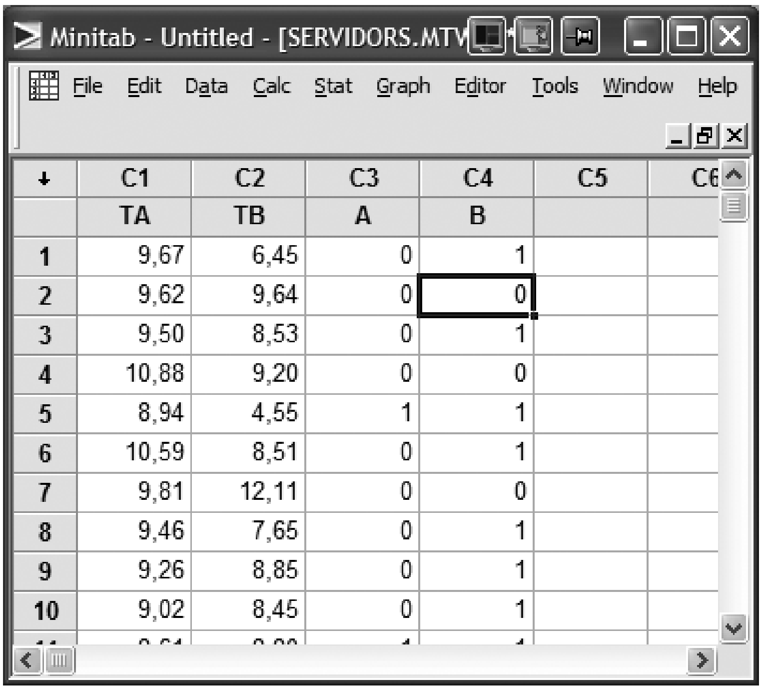


Indicación

Para hacer este ejercicio primero calcularemos una nueva variable, que valga **1** si el tiempo de espera es menor que 9 milisegundos y **0** en caso contrario. Para calcular esta variable, podemos utilizar la instrucción **IF** de Minitab.

Hacemos lo mismo para el tiempo del servidor *B*, creamos una columna de nombre *B* con 1 si el tiempo es inferior a 9 y 0 en caso contrario.

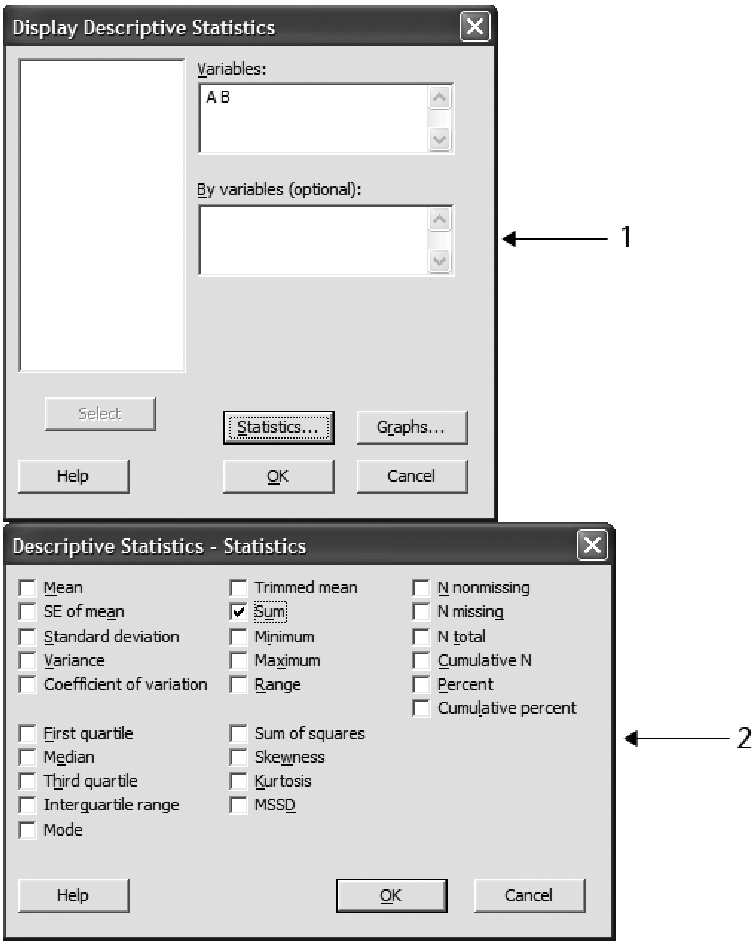
Figura 15. Datos



↓	C1	C2	C3	C4	C5	C6
	TA	TB	A	B		
1	9,67	6,45	0	1		
2	9,62	9,64	0	0		
3	9,50	8,53	0	1		
4	10,88	9,20	0	0		
5	8,94	4,55	1	1		
6	10,59	8,51	0	1		
7	9,81	12,11	0	0		
8	9,46	7,65	0	1		
9	9,26	8,85	0	1		
10	9,02	8,45	0	1		

Una vez tenemos estas dos nuevas columnas, calculamos la suma de cada una de ellas y así tendremos para cada servidor el número de observaciones de tiempo inferior a 9:

Figura 16. Pasos a seguir para obtener el valor suma



Pasos a seguir

Se sigue la ruta ***Stat*** > ***Basic Statistics*** > ***Display Descriptive Statistics. (1)***, y se rellenan los campos en la ventana correspondiente.

En el cuadro de dialogo ***Statistic*** se marca ***Sum (2).***

Seleccionad ***OK*** para obtener el contraste.

Figura 17. Resultados

Descriptive Statistics: A; B

Variable	Sum
A	6.0000
B	31.0000

Para el servidor *A* hay seis observaciones con un tiempo de espera inferior a 9 milisegundos y para el servidor *B* el número de observaciones menores de 9 milisegundos es treinta y una.

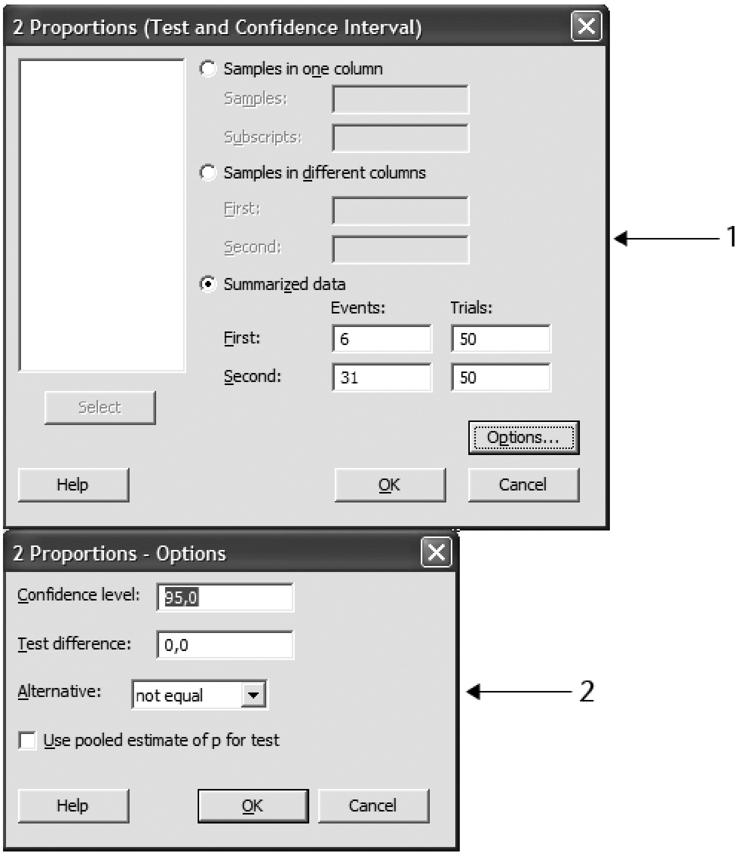
Plantearemos el siguiente contraste:

$$H_0 : p_A - p_B = 0$$
$$H_1 : p_A - p_B \neq 0$$

Fijamos α = 0,05.

La figura 18 muestra los pasos a seguir para realizar el contraste de la diferencia de proporciones.

Figura 18. Pasos para hacer un contraste de hipótesis para la diferencia de proporciones



Pasos a seguir

Se sigue la ruta ***Stat*** > ***Basic Statistics*** > ***2-Proportions*** y se rellenan los campos en la ventana correspondiente.

Summarized data (1)

First: *Events:* 6 *Trials*: 50

Second: *Events*: 31 *Trials*: 50

Seleccione ***Options (2)*** y se rellenan los campos:

Confidence level: 95,0

Alternative: not equal

Los resultados de la figura 19 muestran el *p*-valor = 0,000 < 0,05. Esto indica que podemos rechazar la hipótesis nula y concluimos que la proporción de peticiones con tiempo de espera aceptable es diferente para los dos servidores.

Figura 19. Resultados del contraste de diferencia de proporciones. Minitab

```
Test and CI for Two Proportions

Sample    X   N  Sample p
1         6  50  0,120000
2        31  50  0,620000

Difference = p (1) - p (2)
Estimate for difference:  -0,5
95% CI for difference:  (-0,661908; -0,338092)
Test for difference = 0 (vs not = 0):  Z = -6,05
P-Value = 0,000
```

1.3. Contrastes de hipótesis de comparación de varianzas

Uno de los contrastes desarrollados en el apartado 1.1 para la comparación de medias poblacionales depende del supuesto de igualdad de las dos varianzas poblacionales. Aunque en muchas aplicaciones prácticas este es un supuesto razonable, conviene usar los datos disponibles para contrastar su validez.

En este apartado consideramos el caso de dos muestras aleatorias independientes de poblaciones normales y contrastaremos la igualdad de varianzas poblacionales.

Sea s_1^2 la varianza muestral de una muestra de n_1 observaciones de una población normal con varianza σ_1^2, y s_2^2 la varianza muestral de una muestra independiente de n_2 observaciones de una población normal con varianza σ_2^2. Siempre que las dos varianzas poblacionales sean iguales ($\sigma_1^2 = \sigma_2^2$). La distribución de la relación de las dos varianzas de las muestras s_1^2/s_2^2 está definida por el estadístico F que sigue una **distribución *F* de Snedecor** con $n_1 - 1$ grados de libertad para el numerador y $n_2 - 1$ grados de libertad para el denominador,

$$F_{n_1-1;n_2-1} = \frac{s_1^2}{s_2^2}$$

Contraste de igualdad de varianzas de dos poblaciones normales

Ahora nos interesa contrastar la hipótesis nula que asegura que las varianzas de las poblaciones son iguales $\sigma_1^2 = \sigma_2^2$, es decir, la varianza de la población 1 es igual a la varianza de la población 2. Primero fijaremos el nivel de significación α del contraste.

Nota

A veces, en lugar de:

H_0: $\sigma_1^2 = \sigma_2^2$

escribiremos:

H_0: $\sigma_1^2/\sigma_2^2 = 1$

Hipótesis alternativa, puede ser:

- Bilateral: $\boldsymbol{H_1}$: $\sigma_1^2 \neq \sigma_2^2$, las varianzas de las dos poblaciones son distintas.
- Unilateral: $\boldsymbol{H_1}$: $\sigma_1^2 > \sigma_2^2$, la varianza de la población 1 es mayor que la varianza de la población 2.
- Unilateral: $\boldsymbol{H_1}$: $\sigma_1^2 < \sigma_2^2$, la varianza de la población 1 es menor que la varianza de la población 2.

Bajo el supuesto de la hipótesis nula cierta H_0: $\sigma_1^2 = \sigma_2^2$, el estadístico de contraste es:

$$F^*_{n_1-1;n_2-1} = \frac{s_1^2}{s_2^2}$$

Regla de decisión del contraste de hipótesis

Una vez que se ha calculado el valor del estadístico de contraste, se debe determinar el *p*-valor. El *p*-valor depende de la hipótesis alternativa planteada.

- Si H_1: $\sigma_1^2 \neq \sigma_2^2$, entonces $p\text{-valor} = 2P\left(F_{n_1-1,n_2-1} > F^*\right)$

- Si H_1: $\sigma_1^2 < \sigma_2^2$, entonces p-valor $= P\left(F_{n_1-1,n_2-1} < F^*\right)$
- Si H_1: $\sigma_1^2 > \sigma_2^2$, entonces p-valor $= P\left(F_{n_1-1,n_2 1} > F^*\right)$
- Si p-valor $\leq \alpha$ se rechaza H_0

Ejemplo 4. "Variabilidad de procesadores de texto".

Queremos comparar dos tipos de procesadores de textos: el LaTeX y el OpenOffice. Para hacerlo, consideramos textos más o menos de la misma longitud y contamos la variabilidad del espacio que deja cada procesador entre las palabras. En el caso del LaTeX, consideramos diez textos y obtenemos que la desviación estándar muestral del espacio que deja es de 2,5 mm, mientras que para el OpenOffice consideramos quince textos y obtenemos que la desviación estándar muestral del espacio que deja es de 3,5 mm. Suponiendo normalidad, ¿podemos afirmar que los dos procesadores de textos tienen la misma variabilidad en el espacio que dejan entre palabras?

Para contestar a la pregunta hemos de realizar un contraste de igualdad de varianzas.

El contraste de hipótesis es: $\begin{array}{l} H_0 : \sigma_{LaTeX}^2 = \sigma_{OpenOffice}^2 \\ H_1 : \sigma_{LaTeX}^2 \neq \sigma_{OpenOffice}^2 \end{array}$

Fijamos el valor de $\alpha = 0{,}05$.

El estadístico de contraste vale: $F^* = \dfrac{s_{LaTeX}^2}{s_{OpenOffice}^2}$. Los valores de s_{LaTeX}^2 y $s_{OpenOffice}^2$ son, respectivamente, $s_{LaTeX}^2 = 6{,}25$ y $s_{OpenOffice}^2 = 12{,}25$.

El estadístico F sigue la distribución F de Fisher-Snedecor con 9 y 14 grados de libertad. El valor del estadístico de contraste será:

$$F^* = \frac{6{,}25}{12{,}25} \approx 0{,}51.$$

Los valores críticos serán:

$F_{1-\alpha/2,9,14} = F_{0{,}975,9,14} \simeq 0{,}265$ y $F_{\alpha/2} = F_{0{,}025,9,14} \simeq 3{,}21$.

Para calcular los valores críticos utilizaremos la tabla F o mediante un software estadístico.

Como $F_{0,0{,}25,9,14} < F^* < F_{0{,}975,9,14}$ aceptamos la hipótesis nula y concluimos que las varianzas son iguales. Luego los dos procesadores tienen la misma va-

riabilidad en el espacio que dejan entre palabras. Si quisiéramos realizar el contraste con el *p*-valor, éste valdría: $p = 2 \cdot p(F_{9,14} < 0,51) \approx 0,312$. Como es mucho mayor que 0,05, aceptamos la hipótesis nula y llegamos a la misma conclusión.

En el **ejemplo 1. "Comparación de las medias del tiempo de respuesta de dos servidores"**, cuando realizamos el contraste de diferencia de medias con Minitab **no** presupusimos que las varianzas fueran iguales. Ahora realizaremos un contraste para comparar las dos varianzas y ver si son iguales.

Las hipótesis nula y alternativa son: $\begin{array}{l} H_0 : \sigma_A^2 = \sigma_B^2 \\ H_1 : \sigma_A^2 \neq \sigma_B^2 \end{array}$

Fijamos α = 0,1. El estadístico de contraste es: $F^* = \frac{s_A^2}{s_B^2}$, donde s_A^2 y s_B^2 son respectivamente, las varianzas de los tiempos de espera de los servidores *A* y *B*. La distribución de *F* es la de la *F* de Snedecor con 50 – 1 = 49 grados de libertad en el numerador y 50 – 1 = 49 grados de libertad en el denominador.

Se resolverá el problema con Minitab. Los resultados de Minitab se muestran en la figura 20.

Figura 20. Resultados del contraste de varianzas. Minitab

```
Test for Equal Variances: TA; TB

90% Bonferroni confidence intervals for standard
deviations

     N    Lower     StDev     Upper
TA  50  0,75196  0,90019  1.12176
TB  50  1.45954  1.74726  2.17731

F-Test (Normal Distribution)
Test statistic = 0,27; p-value = 0,000

Levene's Test (Any Continuous Distribution)
Test statistic = 13.14; p-value = 0,000
```

Pasos a seguir

Se sigue la ruta ***Stat*** > ***Basic Statistics*** >***2-Variances*** y se rellenan los campos en la ventana correspondiente.

En el cuadro de dialogo se completan los campos:

Samples in different colums:

First: TA

Second: TB

Seleccionad ***Options***, completad los campos:

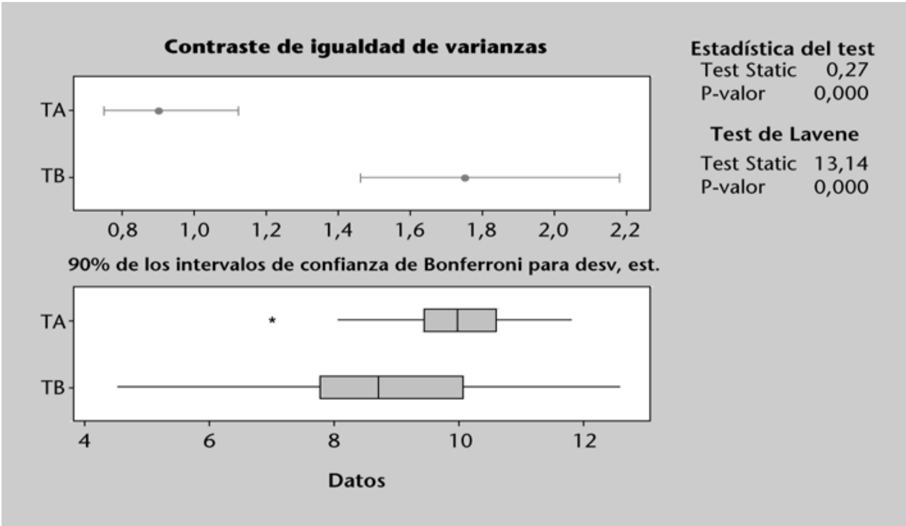
Confidence level: 90,0

Title: Contraste de igualdad de varianzas

Podemos ver que como el *p*-valor es prácticamente cero, hemos de rechazar la hipótesis nula, es decir, no podemos considerar que las varianzas sean iguales. Por esa razón cuando hicimos el contraste de diferencia de medias **no** asumimos que las varianzas fueran iguales.

Minitab también nos ha proporcionado el siguiente gráfico para el contraste de igualdad de varianzas:

Figura 21. Resultados del contraste de igualdad de varianzas. Minitab



La figura 21 presenta un gráfico con los intervalos de confianza de las varianzas de las dos poblaciones, se observa que los intervalos son distintos y no se solapan. El *p*-valor del test *F* indica que se rechaza la hipótesis de igualdad de varianzas.

En el gráfico de *boxplot* se ve claramente que la variabilidad del tiempo de espera del servidor *A* es mucho más pequeña que la del servidor *B*.

2. Comparación de grupos mediante ANOVA

En el apartado anterior se presentaron algunos de los contrastes de hipótesis que se usan habitualmente para determinar si existen diferencias significativas entre dos poblaciones o grupos de individuos. En ocasiones, sin embargo, se deseará comparar más de dos poblaciones o grupos entre sí, para lo cual se emplearán las técnicas de ***analysis of variance*** **(ANOVA)** que se introducen en este apartado.

Nota

El acrónimo **ANOVA** viene del término *analysis of variance* (análisis de la variación existente entre las distintas medias consideradas, para ver si existen diferencias significativas entre las mismas).

Así, por ejemplo, las técnicas ANOVA se podrían aplicar para dar respuestas a preguntas como las siguientes:

- ¿Existen diferencias significativas entre la duración media de los juicios según el tipo de delito cometido (homicidio, abuso sexual, robo, piratería, fraude fiscal, etc.)?

- ¿Existen diferencias significativas entre el gasto anual promedio en tecnología según la franja de edad a la que pertenezca el individuo (niño, joven, adulto, anciano)?

- ¿Existen diferencias significativas entre el número medio de alumnos y ordenadores por centro escolar entre los diferentes países de la eurozona?

- ¿Existen diferencias significativas entre el número medio de autocitas a revistas científicas según la editorial (Elsevier, Inderscience, Taylor & Francis, IGI Global, etc.)?

- ¿Existen diferencias significativas entre el consumo medio de combustible según el modelo de automóvil usado (deportivo, turismo, todoterreno, monovolumen, etc.)?

- ¿Existen diferencias significativas entre la calidad media (medida a partir de unos parámetros definidos) de los resultados de búsquedas en línea según el tipo de motor usado (Google, Microsoft Bing, Yahoo!, etc.)? (figura 22)

Observad

Los ejemplos que se presentan en este capítulo se caracterizan porque la pertenencia a una población o a otra depende de un único factor (tipo de delito, franja de edad, país, editorial, modelo de automóvil, motor de búsqueda, etc.). En estos casos, se usa ANOVA de un único factor (en inglés *one-way* ANOVA o *single-factor* ANOVA). Sin embargo, existen también técnicas ANOVA para el caso en que los grupos vengan determinados por dos factores (p. ej.: tipo de delito y solvencia económica del acusado, franja de edad y clase social, etc.).

Figura 22. ANOVA permite comparar la calidad media de diferentes servicios



2.1. Comparaciones de varias medias

Cuando se desean comparar entre sí las medias correspondientes a más de dos poblaciones o grupos de individuos, se podría pensar en comparar dichas medias dos a dos mediante un contraste de hipótesis para dos poblaciones. Así, por ejemplo, en el caso de tres poblaciones se podría pensar en realizar una serie de tests *t* de hipótesis para comparar las distintas medias entre sí: un primer test *t* para comparar las medias de las poblaciones 1 y 2, otro para comparar las medias de las poblaciones 1 y 3, y otro para comparar las medias de las poblaciones 2 y 3. Sin embargo, esta aproximación tiene un grave problema: si para cada test *t* se usa un nivel de significación α (generalmente se usa $\alpha = 0{,}05$), entonces la probabilidad de cometer un **error de tipo I** es α en cada test; en tales condiciones, se puede comprobar que la probabilidad de cometer un error de tipo I en el global de los tres tests sería de $1-(1-\alpha)^3$ (si $\alpha = 0{,}05$ dicha probabilidad sería de, aproximadamente, 0,14). En otras palabras, comparando las medias dos a dos se está realizando un test global con una probabilidad de error de tipo I mucho mayor que la prevista inicialmente para cada test individual. Para evitar este problema se pueden usar las técnicas ANOVA, que permiten realizar un único test global con una probabilidad de error de tipo I determinada (generalmente $\alpha = 0{,}05$).

Recordad

Un **error de tipo I** consiste en rechazar la hipótesis nula cuando resulta que ésta es cierta. En este caso, la hipótesis nula sería que las medias son coincidentes.

El test *F* de ANOVA

A fin de comparar las medias correspondientes a *k* poblaciones o grupos de individuos distintos ($k \geq 3$), se puede plantear el siguiente contraste de hipótesis, donde el símbolo μ_i representa la media de la población *i*-ésima para $i = 1,2,\ldots,k$:

$$\begin{cases} H_0 : \mu_1 = \mu_2 = \ldots = \mu_k \text{ (todas las medias son iguales)} \\ H_a : \text{no todas las medias son iguales} \end{cases} \qquad (1)$$

En otras palabras, la hipótesis nula, H_0, sostiene que no hay diferencias significativas entre las distintas medias poblacionales, mientras que la hipótesis alternativa, H_a, sostiene todo lo contrario, p. ej.: que las medias sí son significativamente distintas. Es importante observar aquí que la hipótesis nula no dice que todas las medias sean significativamente distintas entre sí, sino simplemente que no todas las medias son iguales, aunque puede haber algunas de ellas que sí lo sean (podría ocurrir, por ejemplo, que $\mu_1 \neq \mu_2$ y $\mu_2 \neq \mu_3$ pero siendo $\mu_1 = \mu_3$). Por tanto, si se concluyese que no todas las medias son iguales, cabría realizar un análisis posterior para determinar cuáles de ellas son diferentes entre sí.

Software estadístico

En la actualidad existe una gran variedad de **programas estadísticos** o de análisis de datos de gran calidad, tanto comerciales (Minitab, SPSS, MS Excel, SAS, S-Plus, etc.) como de código abierto (R, Calc de Open Office, etc.).

El contraste de hipótesis (1) se llama test *F* de ANOVA, y generalmente se recurre al uso de algún ***software* estadístico** para resolverlo, es decir, para obtener el *p*-valor asociado al test. A partir de dicho *p*-valor corresponde al investigador de-

terminar si ha sido posible encontrar suficientes evidencias para rechazar la hipótesis nula o si, por el contrario, los datos empíricos parecen no estar en contradicción con la hipótesis nula y, por tanto, se acepta ésta como válida. Como en cualquier otro tipo de contraste estadístico, antes de resolver el test se suele fijar un valor de significación, α (por lo general $\alpha = 0{,}05$ o bien $\alpha = 0{,}01$). Una vez obtenido el *p*-valor, si $p - valor < \alpha$ se rechaza la hipótesis nula; en caso contrario no hay indicios suficientes para hacerlo y, por tanto, se aceptará la hipótesis nula como válida. La elección del valor concreto para α dependerá del nivel de confianza, $1 - \alpha$, que se desee que tenga la decisión final sobre la aceptación o no de la hipótesis nula. Así, por ejemplo, un $\alpha = 0{,}05$ implicará un nivel de confianza en la decisión final del 95%, mientras que un $\alpha = 0{,}01$ implicará un nivel de confianza en la decisión final del 99%. El problema de seleccionar niveles de confianza excesivamente elevados (superiores al 99% o, lo que es lo mismo, valores de α inferiores a 0,01) es que entonces el contraste de hipótesis se vuelve excesivamente "conservador", de manera que sólo cuando las evidencias empíricas en contra de la hipótesis nula son totalmente abrumadoras (es decir, sólo cuando las diferencias entre algunas de las medias son desproporcionadas), es posible obtener un *p*-valor más pequeño o igual que α. Por ese motivo, en la mayoría de los casos prácticos se suele usar el valor $\alpha = 0{,}05$ o bien $\alpha = 0{,}01$.

Ejemplo de aplicación de ANOVA: comparando el número medio de accesos a contenidos en línea según la posición del enlace en el portal

En un portal web de acceso a publicaciones en línea, se sospecha que la posición que ocupa el enlace a una determinada base de datos afecta al número de consultas diarias que ésta recibe. Para comprobarlo, se han seleccionado al azar un total de 13 días laborables de un mes y, para cada uno de ellos, se ha contabilizado el número de accesos recibidos. La tabla 4 muestra los valores obtenidos, los cuales han sido agrupados según la posición diaria del enlace (en el encabezado de la página, en el margen derecho o en el margen izquierdo).

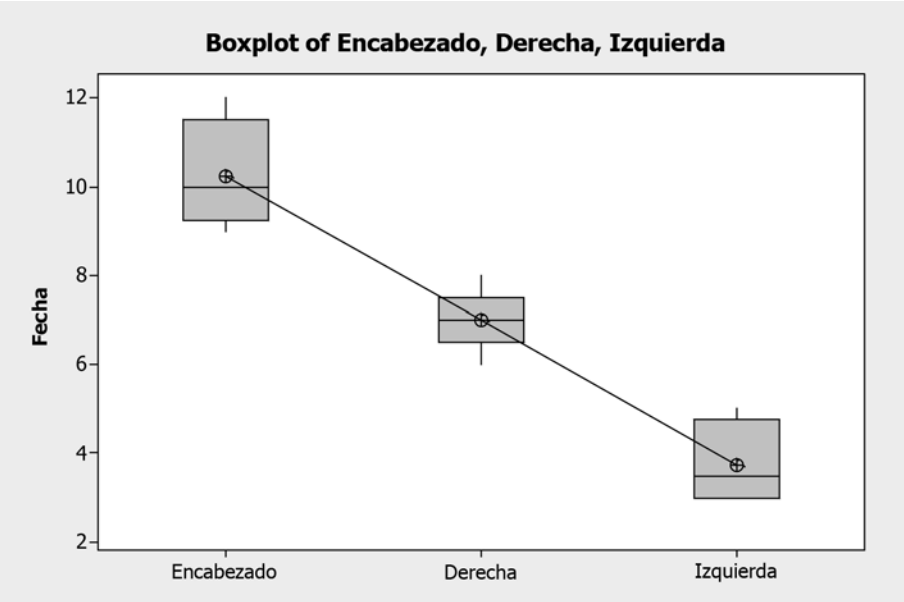
Tabla 4. Accesos a una base de datos según la posición del enlace

	Posición del enlace		
	Encabezado (1)	**Derecha (2)**	**Izquierda (3)**
	10	7	3
	12	6	3
	10	7	5
	9	8	4
		7	
Total	41	45	15
Media	$\bar{x}_1 = 10{,}25$	$\bar{x}_2 = 7{,}0$	$\bar{x}_3 = 3{,}75$

¿Se puede afirmar que hay diferencias significativas entre las distintas medias?, es decir: ¿depende el número medio de consultas diarias de la posición que ocupe el enlace?

Como primera aproximación a este problema, se puede optar por generar un diagrama de cajas y bigotes (*boxplot*) para cada uno de los grupos de datos. La figura 23 muestra dicho diagrama que incluye además una línea uniendo las respectivas medias. Se aprecian claras diferencias entre los tres grupos considerados, tanto a nivel de *boxplots* como a nivel de las respectivas medias.

Figura 23. *Boxplot* del número de consultas para cada posición



Sin embargo, para contestar de forma contundente a las preguntas anteriores, resulta necesario realizar un test F de ANOVA. El contraste de hipótesis se puede formular como sigue:

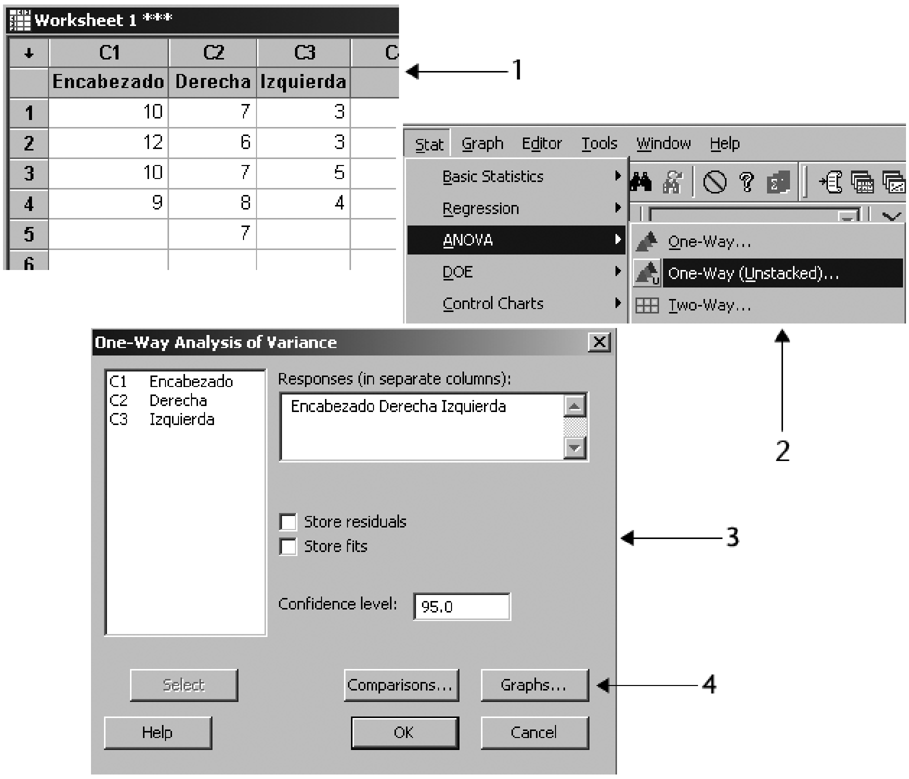
$$\begin{cases} H_0 : \bar{x}_1 = \bar{x}_2 = \bar{x}_3 \\ H_a : \text{no todas las medias son iguales} \end{cases}$$

Para resolver dicho contraste, se fijará un valor de significación $\alpha = 0{,}05$ y se recurrirá al uso de software estadístico para obtener el p-valor correspondiente a las observaciones de la tabla 4.

La figura 24 muestra los pasos básicos necesarios para generar un análisis ANOVA con el **programa Minitab**. Por su parte, la figura 24 muestra el *output* generado para los datos de este ejemplo. Se observa que el valor resultante para el estadístico del contraste es $F = 44{,}47$. El estadístico F es una variable aleatoria que se comporta según una distribución F-Snedecor con 2 grados de libertad en el numerador (*DF Factor*) y 10 grados de libertad en el denominador (*DF Error*). El p-valor no es más que la probabilidad de que una variable aleatoria con esas características supere el valor observado para el estadístico de contraste, p. ej.: p-valor = $P(F_{2,10} > 44{,}47)$. Según se observa en el *output*, en este caso se obtiene p-valor = 0,000. Dado que el p-valor es mucho menor que el nivel de significación escogido (p-valor = $0{,}000 < 0{,}05 = \alpha$), se concluye que los datos obtenidos parecen contradecir la hipótesis nula y, por tanto, ésta se debe rechazar. Así pues, hay indicios claros para pensar que no todas las medias son

iguales, p. ej.: que el número medio de consultas diarias sí depende de la posición que ocupe el enlace.

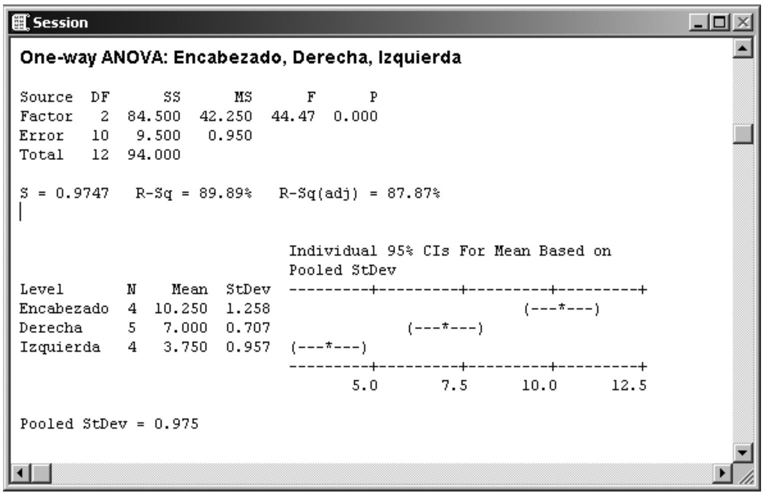
Figura 24. Pasos a seguir para realizar un análisis ANOVA en Minitab



Pasos a seguir

Una vez introducidos los datos en el programa *(1)*, se sigue la ruta ***Stat*** > ***ANOVA*** > ***One-Way (Unstacked)*** *(2)* y se seleccionan las variables en la ventana de ANOVA *(3)*. Más adelante se hará uso de la opción ***Graphs*** de esta ventana *(4)*.

Figura 25. *Output* ANOVA de Minitab para la comparativa de posiciones



En la segunda parte del *output* Minitab se representa cada una de las medias junto con su respectivo intervalo de confianza para un nivel de confianza del 95%. Se observa que los intervalos son disjuntos (no se solapan), lo que significa que las observaciones aportan evidencias de que las tres medias son significativamente distintas. En general, sin embargo, el hecho de que todas las medias no sean iguales no implicará necesariamente que todas sean distintas (es decir, podría haber intervalos que se solapasen y otros que no).

La figura 26 muestra el correspondiente *output* ANOVA que ofrece **Microsoft Excel**. Se observa el mismo valor para el estadístico $F = 44{,}47$, así como un p-valor = 1,0543E-05 (es decir, p-valor = 0,00001543 o, redondeando, p-valor = 0,000).

Figura 26. *Output* ANOVA de Excel para la comparativa de posiciones

	A	B	C	D	E	F	G
1	Análisis de varianza de un factor						
2							
3	RESUMEN						
4	*Grupos*	*Cuenta*	*Suma*	*Promedio*	*Varianza*		
5	Encabezado	4	41	10,25	1,58333333		
6	Derecha	5	35	7	0,5		
7	Izquierda	4	15	3,75	0,91666667		
8							
9							
10	ANÁLISIS DE VARIANZA						
11	*Origen de las variaciones*	*Suma de cuadrados*	*Grados de libertad*	*Promedio de los cuadrados*	*F*	*Probabilidad*	*Valor crítico para F*
12	Entre grupos	84,5	2	42,25	44,4736842	1,0543E-05	4,102821015
13	Dentro de los grupos	9,5	10	0,95			
14							
15	Total	94	12				
16							

Nota

Para poder realizar ANOVA con **MS Excel**, es necesario instalar previamente un complemento llamado "Análisis de datos". Usando Google o cualquier otro buscador es fácil encontrar información detallada sobre el proceso de instalación. También existe un complemento similar para Open Office Calc.

Ejemplo de aplicación de ANOVA: comparando promedios de resultados válidos ofrecidos por un motor de búsqueda según el algoritmo empleado

Los desarrolladores de un nuevo motor de búsqueda especializado en recursos de investigación están probando tres algoritmos distintos de recuperación de la información. Para comprobar si el promedio de resultados válidos que proporciona cada algoritmo es el mismo en los tres casos, se han realizado unas pruebas aleatorias con cada uno de ellos. La tabla 5 muestra las observaciones que se han obtenido tras realizar las pruebas.

Tabla 5. Resultados válidos obtenidos con cada algoritmo

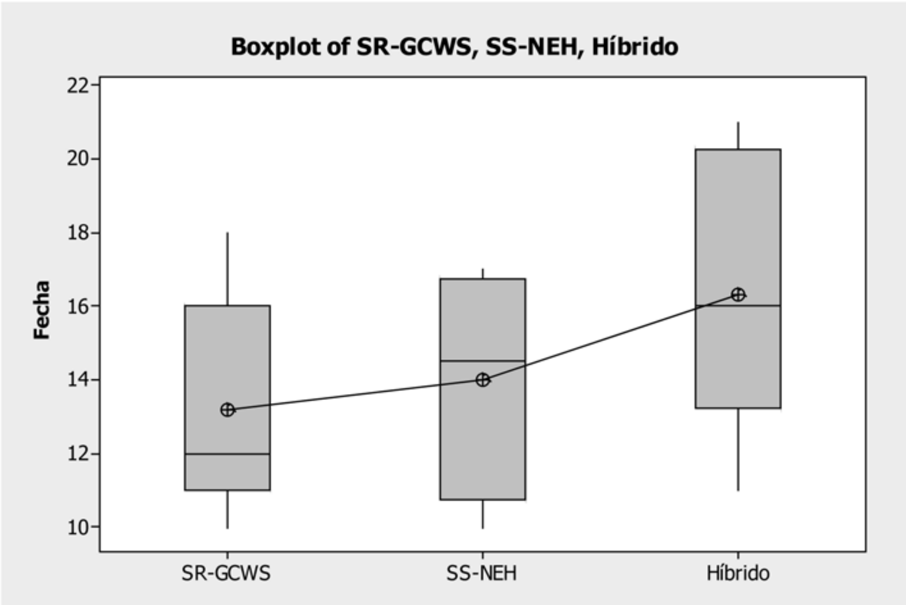
	Algoritmo		
	SR-GCWS	**SS-NEH**	**Híbrido**
	12	10	16
	10	17	14
	18	16	16
	12	13	11
	14		20
			21
Total	66	56	98
Media	$\bar{x}_1 = 13{,}20$	$\bar{x}_2 = 14{,}00$	$\bar{x}_3 = 16{,}33$

¿Se puede afirmar que hay diferencias significativas entre los distintos promedios?, es decir: ¿depende el promedio de resultados válidos obtenidos del algoritmo que implemente el motor de búsqueda?

Nuevamente, para responder adecuadamente a estas preguntas resulta necesario llevar a cabo un test F de ANOVA. Como paso previo, sin embargo, pode-

mos graficar los correspondientes *boxplots*. Como se observa en la figura 27, en este caso las diferencias entre los distintos grupos no parecen ser excesivas, si bien el algoritmo híbrido parece haber proporcionado resultados ligeramente superiores al resto.

Figura 27. *Boxplot* del número de resultados válidos para cada algoritmo



A fin de comprobar si las diferencias entre los promedios son o no estadísticamente significativas, se formula el siguiente contraste ANOVA:

$$\begin{cases} H_0 : \overline{x}_1 = \overline{x}_2 = \overline{x}_3 \\ H_a : \text{no todas las medias son iguales} \end{cases}$$

Nuevamente se hará uso de un nivel de significación $\alpha = 0{,}05$ (es decir, el nivel de confianza usado es del 95%). Las figuras 28 y 29 muestran, respectivamente, los *output* Minitab y Excel para este ejemplo.

Figura 28. *Output* ANOVA de Minitab para la comparativa de algoritmos

```
Session

One-way ANOVA: SR-GCWS, SS-NEH, Híbrido

Source  DF     SS    MS     F      P
Factor   2   29.2  14.6  1.29  0.312
Error   12  136.1  11.3
Total   14  165.3

S = 3.368   R-Sq = 17.66%   R-Sq(adj) = 3.94%


                           Individual 95% CIs For Mean Based on Pooled StDev
Level     N    Mean  StDev    +---------+---------+---------+---------
SR-GCWS   5  13.200  3.033    (------------*------------)
SS-NEH    4  14.000  3.162     (--------------*--------------)
Híbrido   6  16.333  3.724               (-----------*-----------)
                              +---------+---------+---------+---------
                           10.0      12.5      15.0      17.5

Pooled StDev = 3.368
```

Figura 29. *Output* ANOVA de Excel para la comparativa de algoritmos

	A	B	C	D	E	F
1	Análisis de varianza de un factor					
2						
3	RESUMEN					
4	*Grupos*	*Cuenta*	*Suma*	*Promedio*	*Varianza*	
5	SR-GCWS	5	66	13,2	9,2	
6	SS-NEH	4	56	14	10	
7	Híbrido	6	98	16,33333333	13,8666667	
8						
9						
10	ANÁLISIS DE VARIANZA					
11	*Origen de las variaciones*	*Suma de cuadrados*	*Grados de libertad*	*Promedio de los cuadrados*	*F*	*Probabilidad*
12	Entre grupos	29,2	2	14,6	1,28697356	0,311619296
13	Dentro de los grupos	136,1333333	12	11,34444444		
14						
15	Total	165,3333333	14			

En ambos *outputs* se observa un valor del estadístico $F = 1,29$. En esta ocasión, dicho estadístico es una variable aleatoria que se distribuye según una F-Snedecor con 2 grados de libertad en el numerador (*DF Factor*) y 12 en el denominador (*DF Error*). La probabilidad de que una variable como esta alcance o supere el valor 1,29 obtenido por el estadístico es de 0,312, que es precisamente el p-valor que se observa en ambos *outputs*. Puesto que p-valor $= 0,312 > \alpha = 0,05$, no parece que haya indicios suficientes como para rechazar la hipótesis nula. En otras palabras, los datos observados parecen estar en sintonía con la hipótesis nula, por lo que aceptaremos la hipótesis de que los promedios de resultados válidos son equivalentes para los tres algoritmos, sin que haya diferencias estadísticamente significativas entre ellos.

De hecho, en la segunda parte del *output* Minitab se observa que los intervalos de confianza para las tres medias se solapan parcialmente, lo que significa que para un nivel de confianza del 95% no se puede afirmar que haya diferencias significativas entre dichas medias.

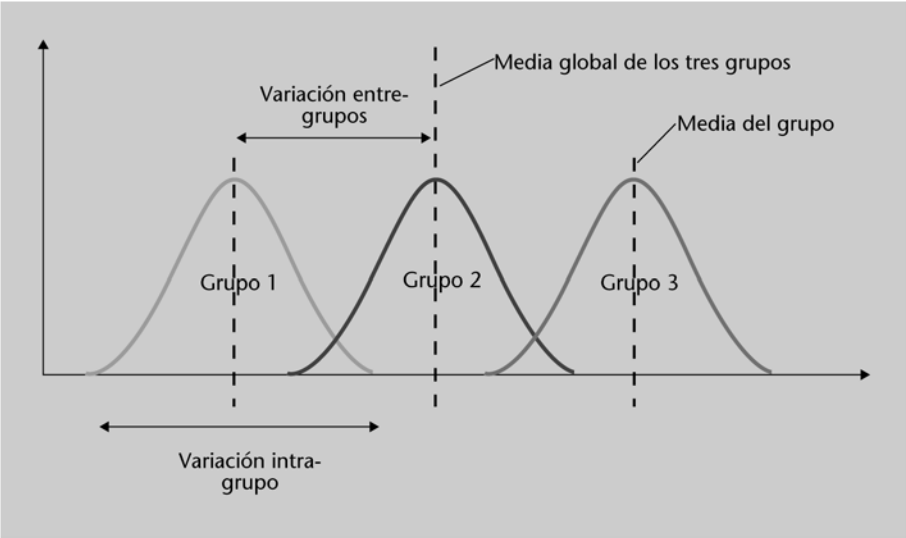
2.2. La lógica del contraste ANOVA

Cuando mediante un experimento aleatorio se recogen una serie de datos (observaciones) y estos son clasificados en varios grupos o niveles según un factor determinado (franja de edad, clase social, etc.), se pueden analizar dos tipos distintos de varianza en las observaciones (figura 30):

- Por un lado, la variación existente entre los distintos grupos o niveles (p.ej.: la variación entre las respectivas medias de cada grupo). Esta se conoce como “variación entre-grupos” o “MS Factor”.

- Por otro, la variación existente dentro de cada grupo o nivel. Esta se conoce como “variación intra-grupos” o “MS Error”.

Figura 30. Variación entre-grupos y variación intra-grupos



En el fondo, lo que hace el test ANOVA es comparar las dos medidas de variabilidad, la variación entre-grupos (MS Factor) y la variación intra-grupos (MS Error). Si ocurre que el MS Factor es significativamente mayor que el MS Error (figura 31), entonces el test concluirá que las medias de los distintos grupos no son iguales en todos los casos (lo que implica que no todos los datos pertenecen a un mismo grupo o, lo que es lo mismo, que el valor de las observaciones sí depende del factor considerado). Si, por el contrario, el MS Factor no es significativamente mayor que el MS Error (figura 32), entonces el test concluirá que no se aprecian diferencias significativas entre las medias de los distintos grupos (en otras palabras, que las observaciones parecen proceder todas de un único grupo o, lo que es lo mismo, que las observaciones no parecen depender del factor considerado).

Figura 31. La variación entre-grupos es mayor que la intra-grupos

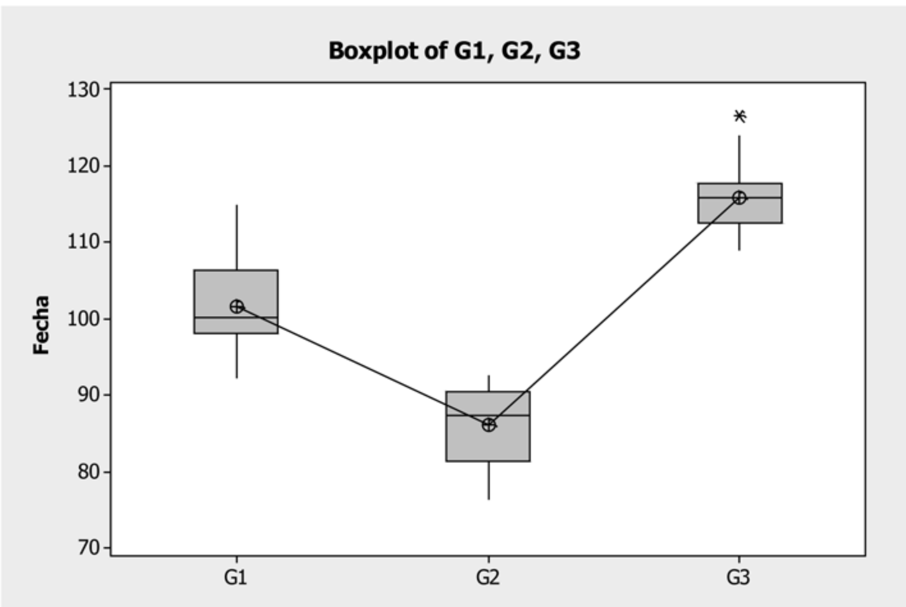
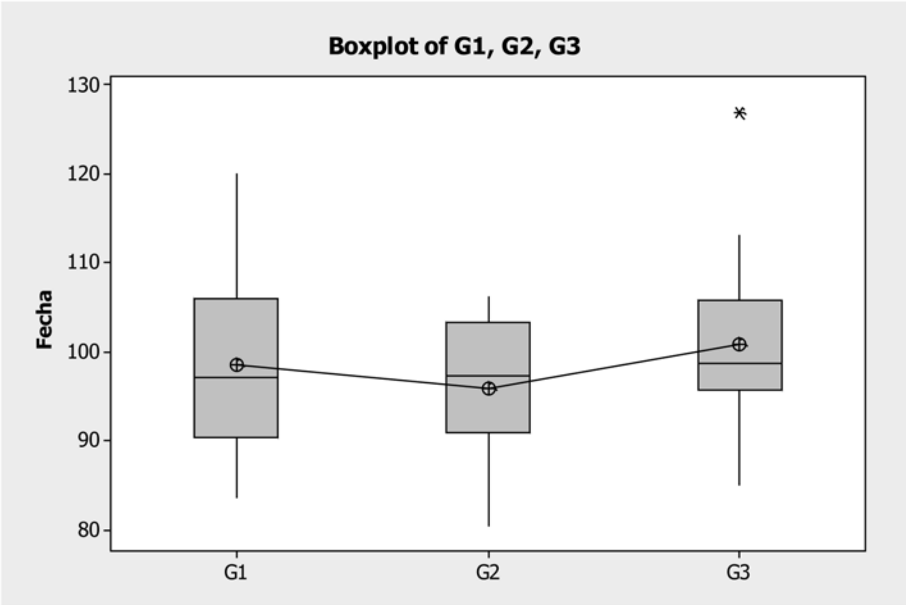


Figura 32. La variación entre-grupos es menor que la intra-grupos



En la figura 25 (*output* ANOVA de Minitab) se observan los valores MS Factor = 42,250 y MS Error = 0,950. Es decir, en este caso la variación entre-grupos (MS Factor) es mucho mayor que la variación intra-grupos (MS Error), lo que ya deja entrever que, probablemente, el test concluya que no todas las medias son iguales. Pero, ¿cómo llega el test a la conclusión final? La figura 33 ayuda a entender mejor cómo funciona el test F de ANOVA:

a) Por un lado, a partir de los valores obtenidos para MS Factor y MS Error se calcula el estadístico de contraste $F = \dfrac{MS\ Factor}{MS\ Error}$.

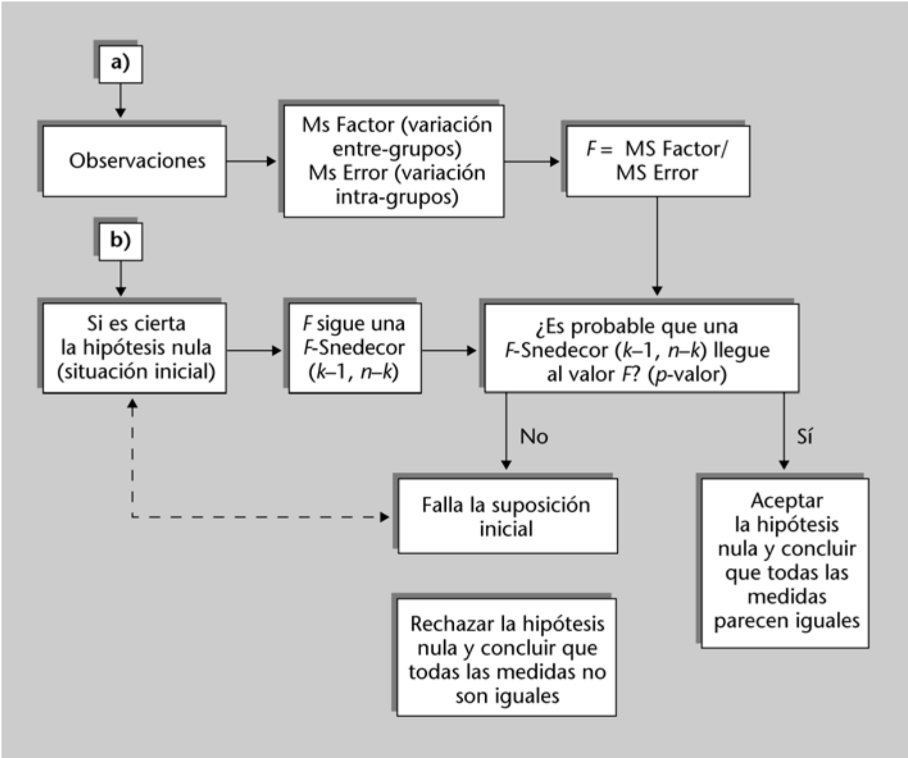
En este caso, F = 44,47.

b) Por otra parte, se sabe que si la hipótesis nula fuese cierta (p. ej.: si todas las medias son iguales), este estadístico F sería una variable aleatoria que seguiría una distribución F-Snedecor con $k - 1$ grados de libertad en el numerador (DF Factor), y $n - k$ grados de libertad en el denominador (DF Error), siendo k el número de grupos o niveles y n el número total de observaciones.

En el ejemplo de la figura 25, DF Factor = 2 y DF Error = 10. Ahora bien, ¿cuál es la probabilidad de que una variable aleatoria F-Snedecor (2, 10) alcance un valor como el obtenido por el estadístico de contraste F? En otras palabras, ¿es razonable pensar que una F-Snedecor (2,10) haya alcanzado un valor de 44,47? La probabilidad de que esto ocurra nos la proporciona el p-valor. De esta manera, un p-valor "pequeño" (inferior al nivel de significación α) se puede interpretar como una probabilidad demasiado baja de que una F-Snedecor (2, 10) pueda dar el valor obtenido para F, lo que pone en entredicho la suposición inicial de que la hipótesis nula era cierta. Por otra parte, un p-valor "grande" (superior al nivel de significación α) se puede interpretar como una

probabilidad aceptable de que, en efecto, una *F*-Snedecor (2, 10) tome dicho valor y, por tanto, no habría evidencias para dudar de la hipótesis nula.

Figura 33. Funcionamiento interno del test F de ANOVA



Observad

que cuando el valor obtenido para el estadístico *F* a partir de las observaciones no es coherente con lo que cabría esperar de una *F*-Snedecor (k – 1, n – k), entonces lo que está fallando es la suposición inicial de que la hipótesis nula es cierta.

2.3. Las hipótesis del modelo ANOVA

Como cualquier otra técnica de inferencia estadística, el contraste ANOVA se puede usar con garantías, para comparar poblaciones o grupos, sólo si se cumplen unas determinadas condiciones de entorno o supuestos básicos:

1) Las observaciones son independientes entre sí y constituyen, para cada población o grupo, una muestra aleatoria.

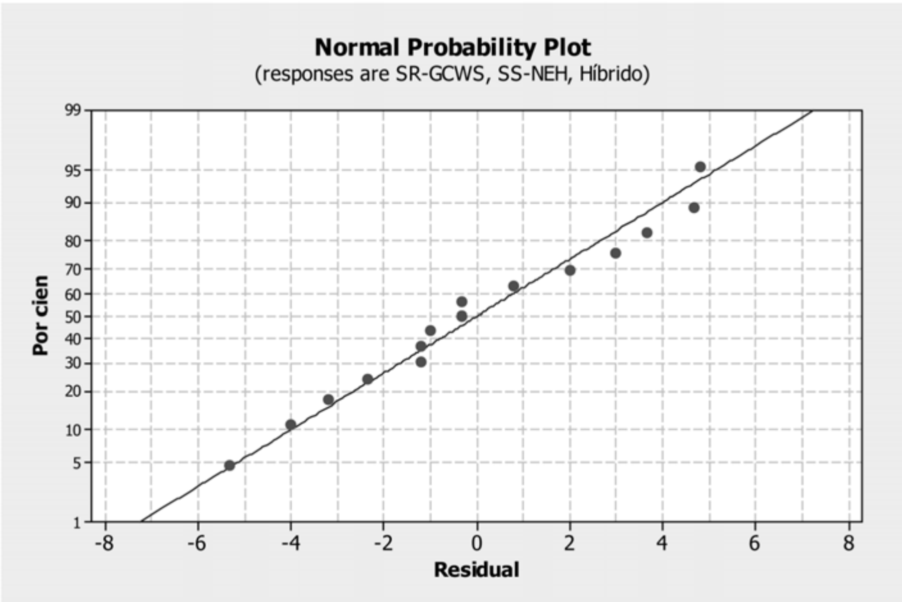
2) Las observaciones de cada población o grupo siguen una distribución aproximadamente normal.

3) Las observaciones de cada población o grupo tienen una varianza σ^2, que es aproximadamente la misma para todos los grupos.

El primer supuesto garantiza que las muestras son aleatorias e independientes, lo que es un requisito común en las técnicas de inferencia estadística. Si las muestras no fuesen aleatorias o las observaciones no fuesen independientes, la información que se generaría estaría sesgada y, por tanto, no sería válida. Es función del investigador garantizar, durante la fase de diseño del experimento y posterior recogida de datos, que se cumple este supuesto.

Por lo que respecta al supuesto segundo (normalidad de los datos), éste se suele comprobar mediante la realización de un gráfico de normalidad para el conjunto de los datos. La figura 34 muestra dicho gráfico para el ejemplo anterior de los algoritmos. Siempre que los puntos (que representan a las observaciones) estén razonablemente cerca de la línea recta (que representa a la distribución normal) y no muestren un patrón de comportamiento extraño, no hay motivos para sospechar que falla el supuesto de normalidad. Si se observase algún patrón de comportamiento anómalo (e.j.: muchos puntos excesivamente alejados de la línea o bien muchos puntos consecutivos situados al mismo lado de la línea), entonces el supuesto de normalidad quedaría en entredicho. Para el ejemplo de los algoritmos, no se observa en el gráfico nada extraño y, por tanto, se puede validar el supuesto de normalidad de los datos.

Figura 34. Gráfico de normalidad para los datos del ejemplo de algoritmos



Pasos a seguir

Este tipo de gráfico se puede obtener con Minitab sin más que marcar la casilla "Normal plot of residuals" en las opciones de Graphs de la ventana ANOVA (figura 24).

Finalmente, por lo que respecta al supuesto de varianza constante, este se suele comprobar o bien calculando las desviaciones estándar de las muestras para verificar que no hay grandes diferencias entre ellas (figura 35), o bien mediante un gráfico que permita comparar visualmente la dispersión de los datos en cada grupo (figura 36). En el caso del ejemplo de los algoritmos no se observan diferencias sustanciales entre las varianzas de los distintos grupos, lo que permite validar el supuesto de varianza constante.

Figura 35. La columna StDev permite estimar la varianza de cada grupo

Session

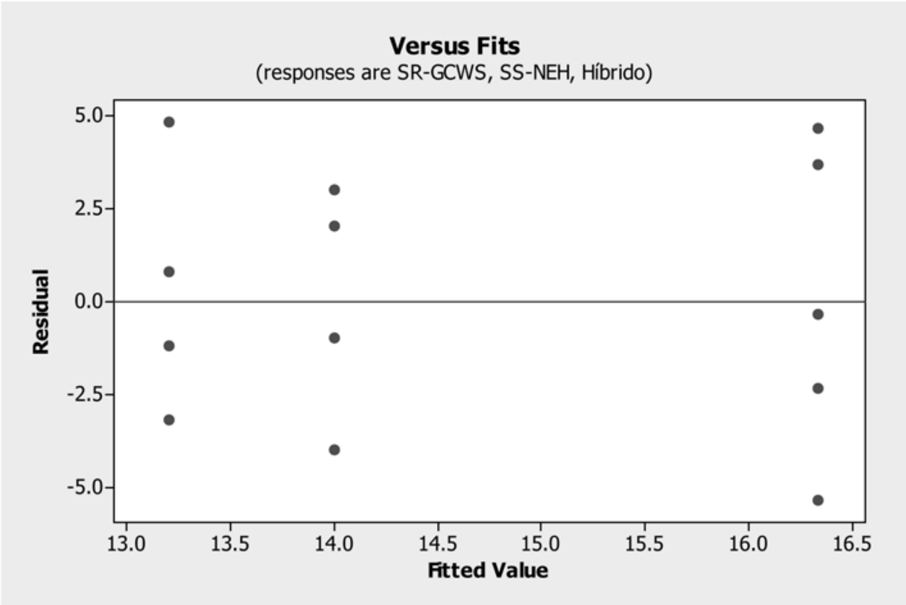
Descriptive Statistics: SR-GCWS, SS-NEH, Híbrido

Variable	N	N*	Mean	SE Mean	StDev	Minimum	Q1	Median	Q3	Maximum
SR-GCWS	5	0	13.20	1.36	3.03	10.00	11.00	12.00	16.00	18.00
SS-NEH	4	0	14.00	1.58	3.16	10.00	10.75	14.50	16.75	17.00
Híbrido	6	0	16.33	1.52	3.72	11.00	13.25	16.00	20.25	21.00

Recordad

La varianza, σ^2, es el cuadrado de la desviación estándar o típica, σ. Por lo general, el valor exacto de la varianza poblacional, σ^2, será desconocido, pero dicho valor se puede estimar mediante la varianza de la muestra, s^2.

Figura 36. El gráfico muestra la dispersión de cada grupo



Pasos a seguir

Este tipo de gráfico se puede obtener con Minitab sin más que marcar la casilla "Residuals versus fits" en las opciones de Graphs de la ventana ANOVA (figura 24).

Ejemplo de aplicación de ANOVA: comparando valoraciones medias en un cuestionario de escala Likert según el perfil de los encuestados

En una universidad se ha implementado recientemente un nuevo servicio online que facilita el acceso a recursos didácticos complementarios. Se desea conocer la opinión de los estudiantes sobre este nuevo servicio y, en particular, si existen diferencias significativas en la valoración media del servicio según la titulación a la que pertenezca el estudiante. Para ello, un investigador ha seleccionado al azar cinco estudiantes de cada uno de los principales estudios que se ofrecen y les ha pedido que rellenen un cuestionario de evaluación del servicio. El cuestionario usa una escala Likert entre 1 (mínima valoración) y 7 (máxima valoración). Los resultados obtenidos se muestran en la tabla 6.

Tabla 6. Valoraciones obtenidas según perfil del estudiante

Estudios				
CC. Información	CC. Empresariales	Ing. Informática	Derecho	Psicología
6	4	5	4	3
5	4	4	4	3
5	3	4	5	2
7	3	6	4	4
4	2	2	6	2

La figura 37 muestra el *output* Minitab correspondiente a los estadísticos descriptivos para cada grupo o nivel de observaciones. A simple vista parecen apreciarse diferencias considerables entre la máxima valoración media (CC. Información, con 5,4) y la mínima (Psicología, con 2,8). El *boxplot* de la figura 38 también apunta a la posibilidad de que las valoraciones medias del servicio

puedan depender del perfil del estudiante, no siendo las mismas para todas las titulaciones.

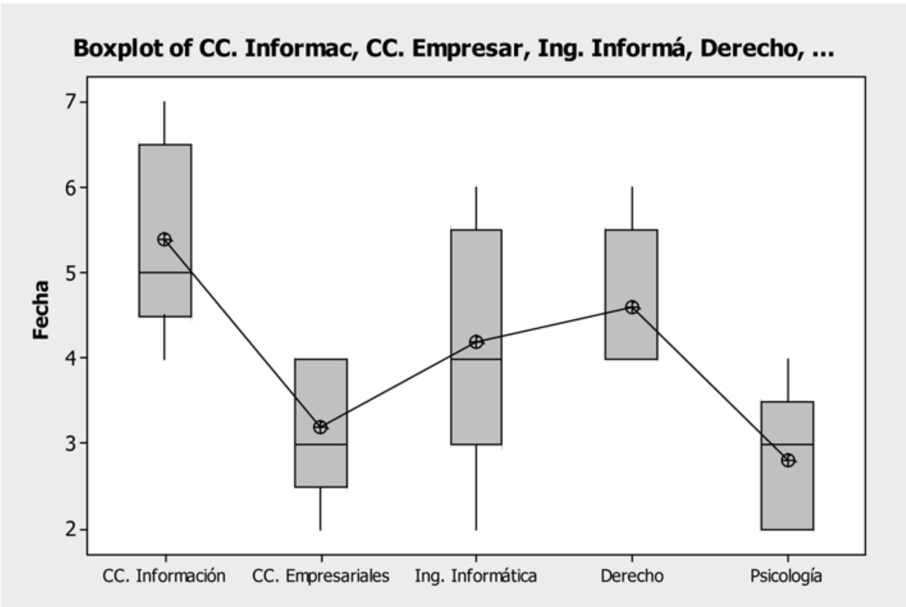
Figura 37. Estadísticos descriptivos de las valoraciones por grupo

Session

Descriptive Statistics: CC. Informac, CC. Empresar, Ing. Informá, Derecho, ...

Variable	N	N*	Mean	SE Mean	StDev	Sum	Minimum	Q1	Median
CC. Información	5	0	5.400	0.510	1.140	27.000	4.000	4.500	5.000
CC. Empresariales	5	0	3.200	0.374	0.837	16.000	2.000	2.500	3.000
Ing. Informática	5	0	4.200	0.663	1.483	21.000	2.000	3.000	4.000
Derecho	5	0	4.600	0.400	0.894	23.000	4.000	4.000	4.000
Psicología	5	0	2.800	0.374	0.837	14.000	2.000	2.000	3.000

Figura 38. *Boxplot* para las valoraciones del servicio por titulación



Para poder corroborar o desmentir estas impresiones visuales de una forma más científica, se opta por realizar un test *F* de ANOVA con un nivel de significación $\alpha = 0{,}01$ (es decir, en este caso se opta por usar un nivel de confianza del 99%).

La figura 39 muestra el *output* ANOVA de Minitab, en el que se aprecia un MS Factor = 5,54 (variación entre-grupos), un MS Error = 1,14 (variación intra-grupos) y un valor para el estadístico de contraste $F = 5{,}54 / 1{,}14 = 4{,}86$. En el supuesto de que la hipótesis nula fuese cierta, este estadístico seguiría una distribución *F*-Snedecor con 4 grados de libertad en el numerador (DF Factor) y 20 grados de libertad en el denominador (DF Error). La probabilidad de que una variable aleatoria *F*-Snedecor (4, 20) tome un valor igual o superior a 4,86 es 0,007 (*p*-valor). Esta probabilidad es extremadamente baja (más baja que el valor de significación fijado), lo cual pone en entredicho el supuesto inicial de que la hipótesis nula era cierta. En otras palabras: puesto que p-valor $< \alpha$ hay que rechazar la hipótesis nula. Así, pues, según las evidencias empíricas encontradas, se puede afirmar con un 99% de confianza que las valoraciones medias de los grupos no son todas iguales.

Figura 39. *Output* ANOVA de Minitab para la comparativa de valoraciones

```
Session
One-way ANOVA: CC. Informac, CC. Empresar, Ing. Informá, Derecho, Psicología

Source  DF     SS    MS     F      P
Factor   4  22.16  5.54  4.86  0.007
Error   20  22.80  1.14
Total   24  44.96

S = 1.068   R-Sq = 49.29%   R-Sq(adj) = 39.15%


Level                N   Mean  StDev
CC. Información      5  5.400  1.140
CC. Empresariales    5  3.200  0.837
Ing. Informática     5  4.200  1.483
Derecho              5  4.600  0.894
Psicología           5  2.800  0.837

                     Individual 99% CIs For Mean Based on Pooled StDev
Level                  +---------+---------+---------+---------
CC. Información                          (--------*--------)
CC. Empresariales         (--------*--------)
Ing. Informática                 (--------*--------)
Derecho                             (--------*--------)
Psicología              (--------*--------)
                       +---------+---------+---------+---------
                     1.5       3.0       4.5       6.0

Pooled StDev = 1.068
```

La segunda parte del *output* Minitab ofrece los intervalos de confianza, a un nivel de confianza del 99% en este caso, para cada una de las medias. Se observa cómo los intervalos más extremos, p. ej.: los correspondientes a CC. Información y Psicología, no se solapan por muy poco. Esto es lógico, puesto que el *p*-valor = 0,007 está muy cercano al valor de significación escogido α = 0,01. Si el *p*-valor hubiera sido todavía menor, ambos intervalos estarían claramente separados. Si, por el contrario, el *p*-valor hubiera sido mayor, ambos intervalos se solaparían parcialmente como ocurre en el resto de los casos.

Antes de dar por definitivas las conclusiones anteriores, conviene validar que se cumplen los supuestos básicos de normalidad y varianza constante de los datos. La figura 40 muestra el gráfico de normalidad correspondiente a las observaciones. No parecen observarse patrones extraños ni demasiados puntos excesivamente alejados de la recta, por lo que se aceptará como válido el supuesto de normalidad. Por su parte, la figura 41 muestra el gráfico de dispersión de cada grupo. Tampoco se observan grandes diferencias entre las dispersiones de los distintos niveles, por lo que se aceptará como válido el supuesto de varianza constante entre los distintos grupos.

Atención

Para evitar confusiones en la segunda parte del *output*, es importante fijar bien el nivel de confianza (1 – α) en la ventana ANOVA (figura 24), de manera que éste se corresponda con el nivel de significación α escogido en cada caso.

Figura 40. Gráfico de normalidad de las valoraciones registradas

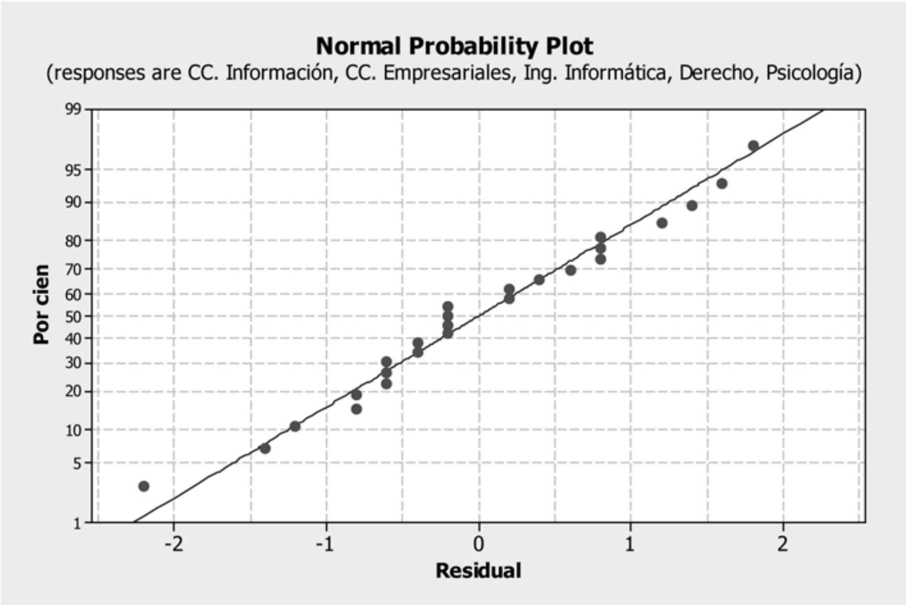
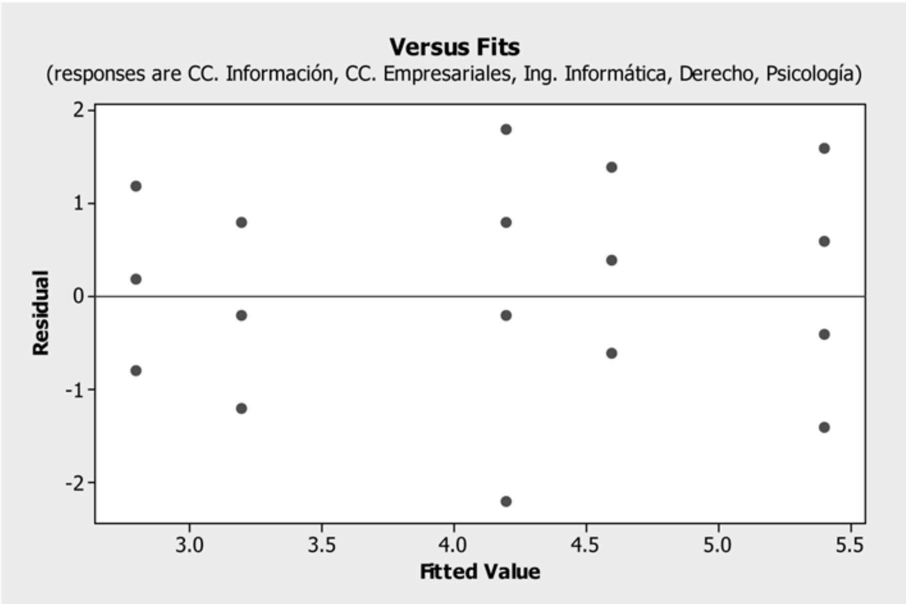


Figura 41. Gráfico de dispersión de cada grupo



Resumen

En este módulo se han presentado las principales técnicas estadísticas que permiten comparar estadísticamente dos o más grupos y discernir si existen o no diferencias significativas entre ellos. En el caso de dos grupos, se usa un contraste de hipótesis basado en la *t*-Student (si se están comparando dos medias) o en la normal (si se están comparando dos proporciones).Por último, se han estudiado procedimientos que se pueden aplicar para hacer inferencias acerca de varianzas poblacionales. Se presentó la distribución *F*, para emplearla en pruebas de hipótesis acerca de las varianzas de dos poblaciones normales.

En el caso de tres o más grupos, se usa un test ANOVA basado en la *F*-Snedecor.

Conviene tener siempre muy presente que lo más importante de un test de hipótesis no son los cálculos matemáticos que subyacen al mismo (en gran parte porque dichos cálculos se pueden automatizar mediante el uso de software), sino la correcta interpretación de los resultados obtenidos y la credibilidad de los mismos, que dependerá de que se cumplan o no los supuestos necesarios para poder aplicar cada una de las técnicas de inferencia vistas en este módulo. Si bien el ordenador puede ser muy útil efectuando los cálculos matemáticos con precisión y rapidez, es responsabilidad del investigador saber interpretar los resultados y comprobar la validez de los supuestos.

Ejercicios de autoevaluación

1) Se estudia el impacto que causa la reubicación forzada sobre la buena vecindad. Se entrevista a seis individuos tanto antes como después de que se les obligara a mudarse. Las entrevistas producen las siguientes puntuaciones:

Entrevistado	Antes	Después
1	2	1
2	1	2
3	3	1
4	3	1
5	1	2
6	4	1

Realizad un contraste de hipótesis al nivel de confianza del 95%.

2) De una muestra de ochenta y cinco mensajes de correo con virus que llegan al servidor de nuestra empresa, nuestro programa KILLVIRUS instalado en el servidor sólo ha detectado veinticinco. Las especificaciones del programa decían que el programa detectaba más del 40% del correo con virus. ¿Estáis de acuerdo con los resultados obtenidos con las especificaciones del programa? (considerad $\alpha = 0{,}1$). Hallad el p-valor del contraste.

3) Queremos comparar la eficiencia de dos compiladores de dos sistemas de indización diferentes: A y B. Para hacerlo, se diseñan ocho programas en cada uno de los dos sistemas y se mide el tiempo de ejecución que tarda cada uno de los programas para resolver ocho problemas determinados de optimización. Los resultados se muestran en la tabla siguiente:

Problema de optimización a resolver	Tiempo de ejecución usado por el ejecutable compilado con el sistema A (en segundos)	Tiempo de ejecución usado por el ejecutable compilado con el sistema B (en segundos)
P1	1,2	1,4
P2	1,3	1,7
P3	1,5	1,5
P4	1,4	1,3
P5	1,7	2,0
P6	1,8	2,1
P7	1,4	1,7
P8	1,3	1,6

¿Podemos asegurar a partir de los datos anteriores que el compilador del sistema A es más eficiente que el compilador del sistema B? (considerad $\alpha = 0{,}05$). Hallad el p-valor del contraste.

4) Dos empresas, A y B, quieren comprar un dispositivo de almacenamiento para realizar copias de seguridad. Antes de hacer la compra se hace un estudio de cuántos gigas necesitarían para realizar la copia. Este estudio consiste en calcular durante diez días toda la información de la empresa necesaria para la copia de seguridad. Los resultados se muestran en la tabla siguiente:

Día	1	2	3	4	5	6	7	8	9	10
Empresa A (gigas)	34	45	47	49	31	30	24	33	35	40
Empresa B (gigas)	45	47	50	42	40	51	46	59	42	46

Suponiendo normalidad y un nivel de significación del 0,05, ¿podemos afirmar que la empresa B necesita más capacidad de almacenamiento que la empresa A? Indicación: antes de nada, tenéis que realizar el contraste correspondiente para ver si las varianzas de las dos muestras son iguales a un nivel de significación de 0,05.

5) Se ha diseñado un experimento aleatorio para analizar durante cuánto tiempo es efectiva cada una de las cuatro drogas distintas que se pueden emplear para aliviar el dolor tras una operación quirúrgica. Los datos obtenidos se muestran en la tabla siguiente:

	Droga			
	A	***B***	***C***	***D***
Tiempo (horas)	8	6	8	4
	6	6	10	4
	4	4	10	2
	2	4	10	
			12	

Para un nivel de significación $\alpha = 005$, contrastar la hipótesis nula de que las cuatro drogas son igualmente efectivas.

6) A la hora de descargar programas *open-source* de Internet, suele ser habitual poder optar por hacerlo desde varios servidores (*mirrors*). Generalmente, las velocidades de descarga desde cada servidor dependen de la distancia existente entre el servidor y el cliente que solicita la descarga. En este caso se desea estudiar si las velocidades de descarga desde cinco servidores distintos se pueden considerar equivalentes o no. Para cada uno de los servidores, se han seleccionado algunos ficheros al azar (todos ellos del mismo tamaño) y se han descargado en el cliente, obteniendo los tiempos de descarga (en segundos) que se muestran en la tabla siguiente:

Servidor					
	A	***B***	***C***	***D***	***E***
Tiempo de descarga (en segundos)	3,8	6,8	4,4	6,5	6,2
	4,2	7,1	4,1	6,4	4,5
	4,1	6,7	3,9	6,2	5,3
	4,4		4,5		5,8

¿Se puede afirmar que la velocidad media de descarga es independiente del servidor seleccionado? Usar un nivel de significación $\alpha = 0{,}01$.

7) Se desean comparar los ingresos por familia (en miles de euros) correspondientes a tres provincias de una misma comunidad autónoma. A tal efecto, para cada provincia se han seleccionado 9 familias al azar y se han registrado sus ingresos. La tabla siguiente muestra las observaciones obtenidas:

Provincia			
	A	***B***	***C***
Ingresos familiares (miles de euros)	45	32	40
	39,5	30	42
	42	37	45
	35	35	39,5
	40	28,5	40
	37	37,5	38
	44	31	51
	48,5	37,6	47,5
	50	25	41

Para un nivel de significación α = 0,05, ¿se puede afirmar que los ingresos medios por familia no dependen de la provincia a la que ésta pertenezca?

8) Una universidad hace uso de tres consultorías externas que ofrecen servicios de asesoramiento técnico en línea a sus estudiantes. Para cada una de estas consultorías, se han escogido al azar seis servicios prestados durante el año en curso y se ha registrado el cambio porcentual en su precio con respecto al precio medio del año anterior. Los datos recogidos se muestran en la tabla siguiente:

	Consultoría		
	A	***B***	***C***
Cambio porcentual en el precio del servicio	3,0	4,5	1,0
	2,5	2,5	–2,5
	–1,5	7,0	–3,5
	4,0	9,0	2,0
	–1,0	1,5	4,6
	5,5	2,0	0,5

Para un nivel de significación α = 0,01, se desea contrastar la hipótesis nula de que el cambio porcentual medio en el precio del servicio es el mismo para las tres consultorías.

9) Se desea comparar el nivel de innovación/originalidad de seis revistas distintas, aunque todas ellas pertenecientes a un mismo ámbito temático. A tal efecto, se han seleccionado al azar siete ejemplares de cada una de las revistas y un comité de expertos ha evaluado el nivel de innovación/originalidad de cada ejemplar, para lo cual se ha usado una escala entre 1 (mínimo) y 300 (máximo). Los datos recogidos se muestran en la tabla siguiente:

	Revista					
	A	***B***	***C***	***D***	***E***	***F***
Nivel de innovación/ originalidad	300	190	228	276	162	264
	300	164	300	296	175	168
	300	238	268	62	157	254
	260	200	280	300	262	216
	300	221	300	230	200	257
	261	132	300	175	256	183
	300	156	300	211	92	93

A partir de estas observaciones, ¿se puede afirmar que todas las revistas muestran un nivel de innovación/originalidad equivalente o, por el contrario, existen diferencias significativas entre los niveles de innovación/originalidad de las distintas revistas? Utilizar un nivel de significación α = 0,05.

Solucionario

1) Se trata de un contraste de diferencia de medias para **muestras dependientes o emparejadas (e**stadístico de contraste t^* =1,49; el valor crítico es $t_{\alpha/2=0,05/2,}$ y 5 grados de libertad = 2,571. $t^* < t_{0,025;\,5}$ no se puede rechazar H_0. Podemos decir con un 95% de confianza que la buena vecindad no ha variado cuando se produce la reubicación.

2) Se trata de un contraste de **diferencia de proporciones**. El estadístico de contraste sigue aproximadamente la distribución normal $N(0,1)$ si el tamaño de la muestra es suficientemente grande como en nuestro caso. El valor del estadístico de contraste es $z^* = -1{,}993$.

El valor crítico será: $z_{0,1} \approx 1{,}28$. Como $z < z_{0,1}$, aceptamos la hipótesis nula y concluimos que las especificaciones del servidor son falsas.

3) Se trata de un contraste de **diferencia de medias dependientes**. El valor del estadístico de contraste vale: $t \approx -3{,}481$. El valor crítico vale $t_{0,05,7} \approx 1{,}895$. Como $t < -t_{0,05;7}$; rechazamos la hipótesis.

El p-valor es $p = p(t_7 < -3{,}481) \approx 0{,}0051$, valor que es menor que 0,05. Por tanto, llegamos a la misma conclusión: rechazar la hipótesis nula.

4) El resultado del Minitab para el **contraste de varianzas** es:

```
Test for Equal Variances: A; B

95% Bonferroni confidence intervals for standard deviations

    N    Lower    StDev    Upper
A  10  5,33911  8,16224  16,4359
B  10  3,60659  5,51362  11,1025
F-Test (Normal Distribution)
Test statistic = 2,19; p-value = 0,258
Levene's Test (Any Continuous Distribution)
Test statistic = 1,61; p-value = 0,221
```

Se acepta la igualdad de varianzas.

El resultado del Minitab para el **contraste de diferencia medias independientes** es:

```
Two-Sample T-Test and CI: Empresa A; Empresa B

Two-sample T for Empresa A vs Empresa B

            N   Mean  StDev  SE Mean
Empresa A  10  36,80   8,16      2,6
Empresa B  10  46,80   5,51      1,7


Difference = mu (Empresa A) - mu (Empresa B)
Estimate for difference:  -10,00
95% upper bound for difference:  -4,60
T-Test of difference = 0 (vs <): T-Value = -3,21   P-Value =
0,002  DF = 18
Both use Pooled StDev = 6,9650
```

Como p-valor < 0,05, rechazamos la hipótesis nula, por lo tanto aceptamos que la empresa B necesita más capacidad de almacenamiento que la empresa A.

5) Estadístico de contraste $F = 12{,}50$; p-valor $= 0{,}001 < \alpha = 0{,}05$ → Rechazar la hipótesis nula, p. ej.: no todos los grupos tienen el mismo comportamiento.

6) Estadístico de contraste $F = 31{,}6$; p-valor = 0,000 < $\alpha = 0{,}01$ → Rechazar la hipótesis nula, p. ej.: no todos los grupos tienen el mismo comportamiento.

7) Estadístico de contraste $F = 13{,}83$; p-valor = 0,000 < $\alpha = 0{,}05$ → Rechazar la hipótesis nula, p. ej.: no todos los grupos tienen el mismo comportamiento.

8) Estadístico de contraste $F = 2{,}91$; p-valor = 0,085 > $\alpha = 0{,}01$ → No rechazar la hipótesis nula, p. ej.: todos los grupos parecen tener el mismo comportamiento.

9) Estadístico de contraste $F = 5{,}30$; p-valor = 0,001 < $\alpha = 0{,}05$ → Rechazar la hipótesis nula, p. ej.: no todos los grupos se comportan igual.

Relación entre variables: causalidad, correlación y regresión

Correlación entre variables. Modelos de regresión simple (lineal, cuadrática, cúbica). Modelos de regresión múltiple

Blanca de la Fuente

PID_00161061

Universitat Oberta de Catalunya

www.uoc.edu

Índice

Introducción

En este módulo se van a estudiar las relaciones que se pueden presentar entre diferentes variables. En concreto se estudiaran posibles relaciones de dependencia entre las variables para intentar encontrar una expresión que permita estimar una variable en función de otras. Para profundizar en el análisis es necesario determinar la *forma* concreta en que se relacionan y medir su *grado* de asociación.

Así, por ejemplo, el estudio de las relaciones entre variables se puede aplicar para dar respuestas a preguntas y casos como los siguientes:

- ¿Existe relación entre la edad de los lectores y el número de préstamos de libros?

- En otro caso, una editorial podría usar la relación entre el número de páginas de un trabajo y el tiempo de impresión para predecir el tiempo empleado en la impresión.

- Se quiere estudiar el "tiempo de respuesta" de unos ciertos programas de búsqueda bibliográfica en función del "número de instrucciones" en que están programados.

- En una determinada empresa de venta de libros en línea, ¿cómo representamos que el aumento de la cantidad gastada en publicidad provoca un incremento de las ventas?

Este módulo examina la relación entre dos variables, una variable independiente y otra dependiente, por medio de la regresión simple y la correlación. También se considera el modelo de regresión múltiple en el que aparecen dos o más variables independientes.

Objetivos

Los objetivos académicos del presente módulo se describen a continuación:

1. Comprender la relación entre correlación y regresión simple.

2. Usar gráficos para ayudar a comprender una relación de regresión.

3. Ajustar una recta de regresión e interpretar los coeficientes.

4. Obtener e interpretar las correlaciones y su significación estadística.

5. Utilizar los residuos de la regresión para comprobar la validez de las suposiciones necesarias para la inferencia estadística.

6. Aplicar contrastes de hipótesis.

7. Ajustar una ecuación de regresión múltiple e interpretar los resultados.

1. Relación entre variables

Cuando se estudian conjuntamente dos o más variables que no son independientes, la relación entre ellas puede ser **funcional** (relación matemática exacta entre dos variables, por ejemplo, espacio recorrido por un vehículo que circula a velocidad constante y el tiempo empleado en recorrerlo) o **estadística** (no existe una expresión matemática exacta que relacione ambas variables, existe una relación aproximada entre las dos variables, por ejemplo, incremento de las ventas de libros en función de la cantidad gastada en publicidad). En este último caso interesa estudiar el grado de dependencia existente entre ambas variables. Lo realizaremos mediante el **análisis de correlación** y, finalmente, desarrollaremos un modelo matemático para estimar el valor de una variable basándonos en el valor de otra, en lo que llamaremos **análisis de regresión**.

El análisis de regresión **no se puede** interpretar como un procedimiento para establecer una relación **causa-efecto o causalidad** entre variables. La regresión solo puede indicar cómo están **asociadas** las variables entre sí y nos permite construir un modelo para explicar la relación entre ellas. La correlación indica el grado de la relación entre dos variables sin suponer que una alteración en una cause un cambio en la otra variable.

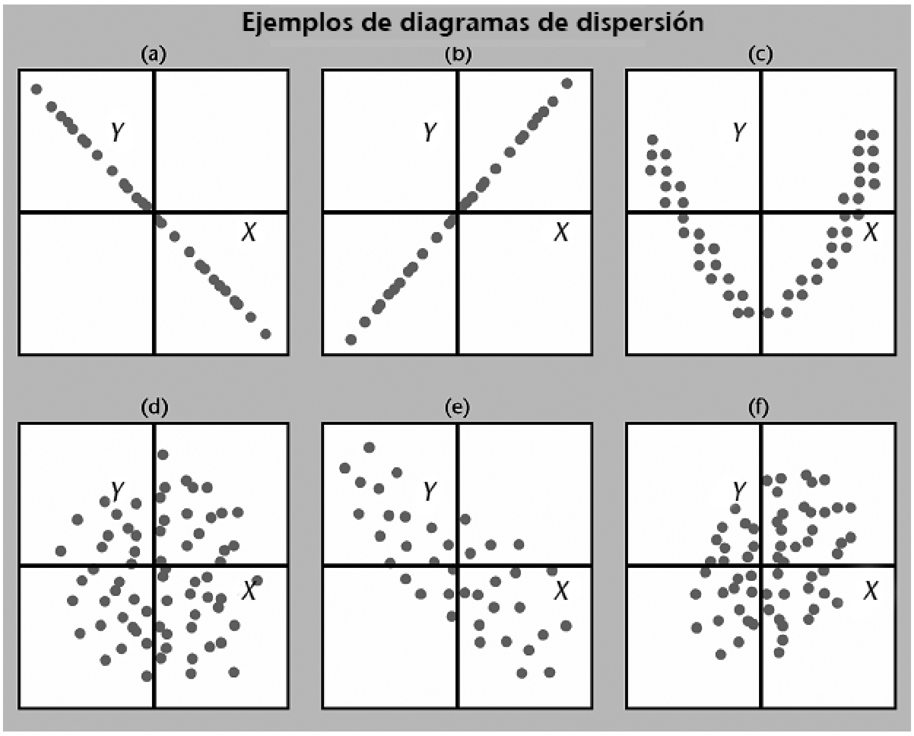
El objetivo principal del análisis de regresión es explicar el comportamiento de una **variable dependiente *Y*** (endógena o explicada) a partir de una o varias **variables independientes** (exógenas o explicativas). El tipo más sencillo de regresión es la **regresión simple**. La regresión lineal simple estima una ecuación lineal que describe la relación, mientras que la correlación mide la fuerza de la relación lineal. Aparte de los modelos lineales se pueden establecer otros modelos de regresión no lineales. El análisis de regresión donde intervienen dos o más variables independientes se llama análisis de regresión múltiple, donde una variable viene explicada por la acción simultánea de otras variables.

Diagrama de dispersión

Antes de abordar el problema, se puede intuir si existe relación entre las variables a través de la representación gráfica llamada **diagrama de dispersión** o **nube de puntos**.

A partir de un conjunto de observaciones (x_i, y_i) de dos variables X e Y sobre una muestra de individuos se representan estos datos sobre un eje de coordenadas x–y. En la figura 1 se incluyen varias gráficas de dispersión que ilustran diversos tipos de relación entre variables.

Figura 1. Diagramas de dispersión



En los casos (a) y (b) tenemos que las observaciones se encuentran sobre una recta. En el primer caso, con pendiente negativa, indica una relación inversa entre las variables (a medida que X aumenta, la Y es cada vez menor) y lo contrario en el segundo caso, en el que la pendiente es positiva, indica una relación directa entre las variables (a medida que aumenta X, la Y también aumenta). En estos dos casos los puntos se ajustan perfectamente sobre la recta, de manera que tenemos una relación funcional entre las dos variables dada por la ecuación de la recta.

En el caso (c) los puntos se encuentran situados en una franja bastante estrecha que tiene una forma bien determinada. No será una relación funcional, ya que los puntos no se sitúan sobre una curva, pero sí que es posible asegurar la existencia de una fuerte relación entre las dos variables. De todos modos, vemos que no se trata de una relación lineal (la nube de puntos tiene forma de parábola).

En el caso (d) no tenemos ningún tipo de relación entre las variables. La nube de puntos no presenta una forma bien determinada; los puntos se encuentran absolutamente dispersos.

En los casos (e) y (f) podemos observar que sí existe algún tipo de relación entre las dos variables. En el caso (e) podemos ver un tipo de dependencia lineal con pendiente negativa, ya que a medida que el valor de X aumenta, el valor de Y disminuye. Los puntos no están sobre una línea recta, pero se acercan bastante, de manera que podemos pensar en una relación lineal. En el caso (f) observamos una relación lineal con pendiente positiva, pero no tan fuerte como la anterior.

Después de estudiar el diagrama de dispersión, el siguiente paso es comprobar analíticamente la dependencia o independencia de ambas variables.

2. Análisis de la correlación

El análisis de correlación mide el grado de relación entre las variables. En este apartado veremos el análisis de correlación simple, que mide la relación entre sólo una variable independiente (X) y la variable dependiente (Y). En el apartado 4 de este módulo se describe el análisis de correlación múltiple que muestra el grado de asociación entre dos o más variables independientes y la variable dependiente.

La correlación simple determina la cantidad de variación conjunta que presentan dos variables aleatorias de una distribución bidimensional. En concreto, cuantifica la dependencia lineal, por lo que recibe el nombre de correlación lineal. El coeficiente de correlación lineal se llama coeficiente de correlación de Pearson designado ***r***, cuyo valor oscila entre –1 y +1. Su expresión es el cociente entre la covarianza muestral entre las variables y el producto de sus respectivas desviaciones típicas:

$$r = \frac{Cov(X,Y)}{S_X S_Y}$$

El valor de r se aproxima a +1 cuando la correlación tiende a ser lineal directa (mayores valores de X significan mayores valores de Y), y se aproxima a –1 cuando la correlación tiende a ser lineal inversa. Podemos formular la pregunta: ¿a partir de qué valor de r podemos decir que la relación entre las variables es fuerte? Una regla razonable es decir que la relación es débil si $0 \leq |r| \leq 0{,}5$; fuerte si $0{,}8 \leq |r| \leq 1$, y moderada si tiene otro valor.

Dada una variable X con $x_1, x_2 \ldots x_n$ valores muestrales y otra variable Y con $y_1, y_2 \ldots y_n$ valores muestrales, siendo n el número total de observaciones y siendo la media de X: $\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}$ y la media de Y: $\overline{y} = \frac{\sum_{i=1}^{n} y_i}{n}$

La covarianza muestral entre dos variables X e Y nos permite medir estas relaciones positivas y negativas entre las variables X e Y:

$$Cov(X,Y) = S_{XY} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})(y_j - \overline{y})$$

La covarianza muestral podemos calcularla mediante otra expresión equivalente:

$$S_{XY} = \frac{\left[\sum_{i,j=1}^{n} x_i y_j\right] - n \cdot \overline{x} \cdot \overline{y}}{n-1}$$

Ejemplo 1. "Estudio de los servicios ofrecidos por un centro de documentación".

Estamos realizando un proceso de evaluación de los servicios ofrecidos por un centro de documentación. Para conocer la opinión de los usuarios se les ha pedido que rellenen un cuestionario de evaluación del servicio. Hacemos dos preguntas, una para que valoren de 0 a 10 su impresión sobre el funcionamiento global del centro y otra pregunta que valora específicamente la atención a los usuarios, para determinar si las valoraciones respecto a la atención al usuario (representadas por la variable dependiente *Y*) están relacionadas con las valoraciones obtenidas respecto al funcionamiento global del centro (variable independiente *X*).

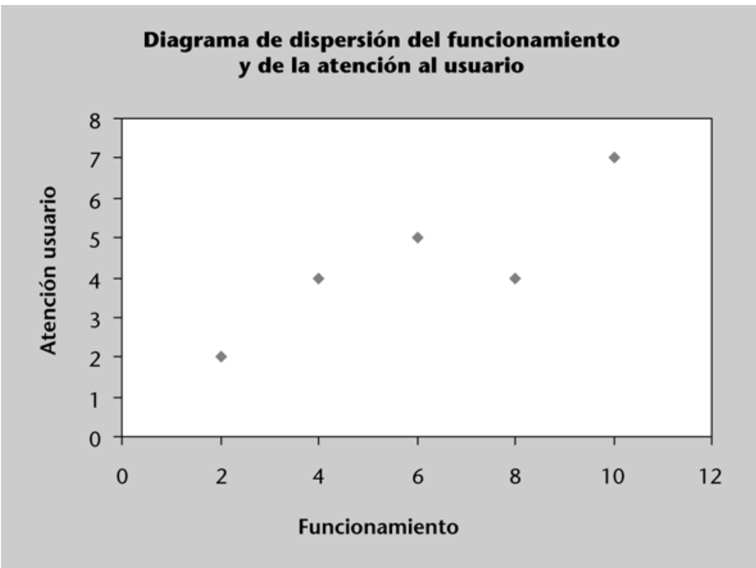
Para ello, un investigador ha seleccionado al azar cinco personas entrevistadas y dan las siguientes valoraciones:

Tabla 1. Datos obtenidos de respuestas a cinco entrevistas realizadas sobre valoraciones de funcionamiento y atención a usuarios de un centro de documentación

Entrevista (*i*)	Funcionamiento (*X*)	Atención (*Y*)
1	2	2
2	4	4
3	6	5
4	8	4
5	10	7

El diagrama de dispersión (figura 2) nos permite observar gráficamente los datos y sacar conclusiones. Parece que las valoraciones de atención al usuario son mejores para valoraciones elevadas del funcionamiento global del centro. Además, para esos datos la relación entre la atención al usuario y el funcionamiento parece poder aproximarse a una línea recta; realmente parece haber una relación lineal positiva entre *X* e *Y*.

Figura 2. Diagrama de dispersión del funcionamiento del centro y de la atención al usuario



Para determinar si existe correlación lineal entre las dos variables, calculamos el coeficiente de correlación ***r***.

En la tabla 2 se desarrollan los cálculos necesarios para determinar los valores de las varianzas, desviaciones típicas muestrales y covarianza muestral.

Tabla 2. Cálculo de las sumas de cuadrados para la ecuación estimada de regresión de mínimos cuadrados

Funcionamiento (*X*)	**Atención (*Y*)**	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$
2	2	–4	–2,4	9,6	16	5,76
4	4	–2	–0,4	0,8	4	0,16
6	5	0	0,6	0	0	0,36
8	4	2	–0,4	–0,8	4	0,16
10	7	4	2,6	10,4	16	6,76

y_i representa las valoraciones observadas (reales) del funcionamiento global obtenidas en la entrevista *i*,

$$n = 5 \quad \sum_{i=1}^{5} x_i = 30 \quad \sum_{i=1}^{5} y_i = 4,4 \quad \sum_{i=1}^{5} (x_i - \bar{x})(y_i - \bar{y}) = 20 \quad \sum_{i=1}^{5} (x_i - \bar{x})^2 = 40 \quad \sum_{i=1}^{5} (y_i - \bar{y})^2 = 13,2$$

realizando las siguientes operaciones obtendremos el coeficiente de correlación lineal.

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{30}{5} = 6 \ ; \ S_X = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}} = \sqrt{\frac{40}{5-1}} = 3,16$$

$$\bar{y} = \frac{\sum_{j=1}^{n} y_j}{n} = \frac{22}{5} = 4,4 \ ; \ S_Y = \sqrt{\frac{\sum_{j=1}^{n} (y_j - \bar{y})^2}{n-1}} = \sqrt{\frac{13,2}{5-1}} = 1,82$$

$$Cov(X,Y) = \frac{1}{n-1} \sum_{i,j=1}^{n} (x_i - \bar{x})(y_j - \bar{y}) = \frac{1}{5-1} \cdot 20 = 5$$

El coeficiente de correlación lineal es:

$$r = \frac{Cov(X,Y)}{S_X S_Y} = \frac{5}{3,16 \cdot 1,82} = 0,87$$

Como el valor del coeficiente de correlación lineal es próximo a 1, se puede afirmar que existe una correlación lineal positiva entre las valoraciones obte-

nidas de atención al usuario y las valoraciones del funcionamiento global del centro. Es decir el, funcionamiento global está asociado positivamente a la atención al usuario.

3. Modelos de regresión simple

3.1. Modelos de regresión lineal simple

Una vez que hemos obtenido el diagrama de dispersión y después de observar una posible relación lineal entre las dos variables, el paso siguiente sería encontrar la ecuación de la recta que mejor se ajuste a la nube de puntos. Esta recta se denomina **recta de regresión.** Una recta queda bien determinada si el valor de su pendiente (b) y de la ordenada en el origen (a) son conocidas. De esta manera la ecuación de la recta viene dada por:

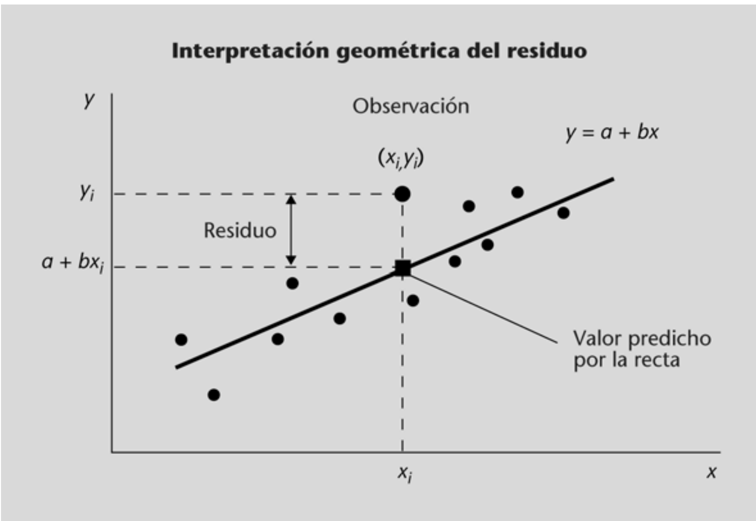
$$Y = a + bx$$

A partir de la fórmula anterior definimos para cada observación (x_i, y_i) el *error* o *residuo* como la distancia vertical entre el punto (x_i, y_i) y la recta, es decir:

$$y_i - (a + bx_i)$$

Por cada recta que consideremos, tendremos una colección diferente de residuos. Buscaremos la recta que minimice la suma de los cuadrados de los residuos. Este es el **método de los mínimos cuadrados**, un procedimiento para encontrar la ecuación de regresión que consiste en buscar los valores de los coeficientes a y b de manera que la suma de los cuadrados de los residuos sea mínima, obteniéndose la **recta de regresión por mínimos cuadrados** (figura 3).

Figura 3. Recta de regresión por mínimos cuadrados



Nota

La recta de regresión pasa por el punto $(\bar{x}, \bar{y})$.

Hemos hecho un cambio en la notación para distinguir de manera clara entre una recta cualquiera: $y = a + bx$ y la recta de regresión por mínimos cuadrados obtenida al determinar a y b.

A partir de ahora, la **recta de regresión** la escribiremos de la manera siguiente:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

El modelo de regresión lineal permite hallar el valor esperado de la variable aleatoria *Y* cuando *X* toma un valor específico.

La **recta de regresión *Y*/*X*** permite predecir un valor de *y* para un determinado valor de *x*.

Para cada observación (x_i,y_i) definimos:

- El valor estimado o predicho para la recta de regresión:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

- Los parámetros o coeficientes de la recta y vienen dados por:

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x} \quad \text{y} \quad \hat{\beta}_1 = \frac{Cov(XY)}{S_X^2} = \frac{S_{XY}}{S_X^2}$$

Siendo:

$\hat{\beta}_0$ es la ordenada en el origen de la ecuación estimada de regresión.

$\hat{\beta}_1$ es la pendiente de la ecuación estimada de regresión.

S_{XY} la covarianza muestral, S_X^2 la varianza muestral de *X*, $\overline{x}$ *e* $\overline{y}$ son las medias aritméticas de las variables *X* e *Y* respectivamente.

- El residuo o error es la diferencia entre el valor observado y_i y el valor estimado $\hat{y}_i$:

$$e_i = y_i - \hat{y}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$$

Ejemplo 1. "Estudio de los servicios ofrecidos por un centro de documentación".

Hemos comprobado en el ejemplo anterior que existe correlación lineal entre ambas variables, ahora calcularemos la **recta de regresión por mínimos cuadrados *Y*/*X*.**

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

en la que,

x_i = valor de funcionamiento para la i-ésima entrevista

$\hat{\beta}_0$ = ordenada en el origen de la línea estimada de regresión

$\hat{\beta}_1$ = pendiente de la línea estimada de regresión

$\hat{y}_i$ = valor estimado de la atención al usuario para la i-ésima entrevista

Para que la línea estimada de regresión ajuste bien con los datos, las diferencias entre los valores observados y los valores estimados de atención al usuario deben ser pequeñas.

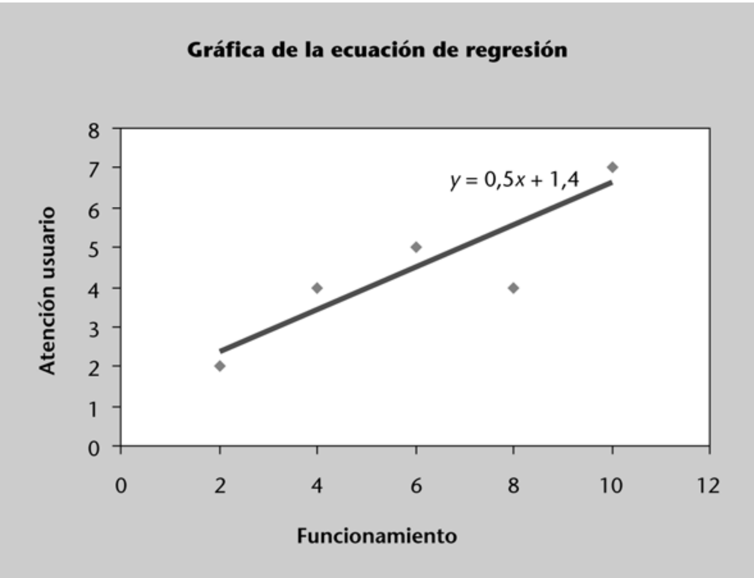
Utilizando los valores obtenidos en la tabla 2 podemos determinar la pendiente y la ordenada en el origen de la ecuación estimada de regresión en este ejemplo. Los cálculos son los siguientes:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{5}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{5}(x_i - \bar{x})^2} = 0,5\ ; \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x} = 1,4$$

Por lo anterior, la ecuación estimada de regresión deducida con el método de mínimos cuadrados, será:

$$\hat{y} = 1,4 + 0,5x$$

Figura 4. Gráfica de la ecuación de regresión ejemplo 1



Interpretación de los parámetros de la recta de regresión

Es importante interpretar los coeficientes de la ecuación en el contexto del fenómeno que se está estudiando.

- Interpretación de la ordenada en el origen, $\hat{\beta}_0$:

  Este coeficiente representa la estimación del valor de *Y* cuando *X* es igual a cero. No siempre tiene una interpretación práctica. Para que sea posible, es preciso que:

  – realmente sea posible que *X* tome el valor $x = 0$,
  – se tengan suficientes observaciones cercanas al valor $x = 0$.

- Interpretación de la pendiente de la recta, $\hat{\beta}_1$:

  Este coeficiente representa la estimación del incremento que experimenta la variable *Y* cuando *X* aumenta en una unidad. Este coeficiente nos informa de cómo están relacionadas las dos variables en qué cantidad varían los valores de *Y* cuando varían los valores de la *X* en una unidad.

La calidad o bondad del ajuste

Una vez acumulada la recta de regresión por mínimos cuadrados debemos analizar si este ajuste al modelo es lo bastante bueno. Mirando si en el diagrama de dispersión los puntos experimentales quedan muy cerca de la recta de regresión obtenida, podemos tener una idea de si la recta se ajusta o no a los datos, pero nos hace falta un valor numérico que nos ayude a precisarlo. La medida de bondad de ajuste para una ecuación de regresión es el **coeficiente de determinación R^2.** Nos indica el grado de ajuste de la recta de regresión a los valores de la muestra y se define como la proporción de varianza en *Y* explicada por la recta de regresión. La expresión de R^2 es la siguiente:

$$R^2 = \frac{\text{Varianza en } Y \text{ explicada por la recta de regresión}}{\text{Varianza total de los datos } Y}$$

La varianza explicada por la recta de regresión es la varianza de los valores estimados y la varianza total de los datos es la varianza de los valores observados. Por tanto, podemos establecer que:

Varianza total de *Y* = varianza explicada por la regresión +
+ varianza no explicada (residual o de los errores)

Es decir, podemos descomponer la variabilidad total (*SSTotal*) de las observaciones de la forma:

$$SSTotal = SSRegresión + SSError$$

en la que,

SSTotal, es la suma de cuadrados totales $SST = \sum_{i=1}^{n}(y_i - \bar{y})^2$

SSRegresión, mide cuánto se desvían los valores de $\hat{y}_i$ medidos en la línea de regresión, de los valores de $\bar{y}_i$, $SSR = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$

SSError, representa el error que se comete al usar $\hat{y}_i$ para estimar y_i, es la suma de cuadrados de estos errores, $SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} e_i^2$

Ahora vemos cómo se pueden utilizar las tres sumas de cuadrados, *SST, SSR* y *SSE* para obtener la medida de bondad de ajuste para la ecuación de regresión, que es el coeficiente de determinación R^2. Vendrá dado por la expresión:

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

- Los valores del coeficiente de determinación están comprendidos entre cero y uno: $0 \leq R^2 \leq 1$
- $R^2 = 1$ cuando el ajuste es perfecto, es decir, todos los puntos están sobre la recta de regresión.
- $R^2 = 0$ muestra la inexistencia de relación entre las variables *X* e *Y*.
- Como R^2 explica la proporción de variabilidad de los datos explicada por el modelo de regresión, cuanto más próximo a la unidad, será mejor el ajuste.

Relación entre R^2 y *r*

Es muy importante tener clara la diferencia entre el coeficiente de correlación y el coeficiente de determinación:

- R^2 mide la proporción de variación de la variable dependiente explicada por la variable independiente.
- r^2 es el coeficiente de correlación, mide el grado de asociación lineal entre las dos variables.
- No obstante, en la regresión lineal simple tenemos que $R^2 = r^2$.

Observaciones

Un coeficiente de determinación diferente de cero no significa que haya relación lineal entre las variables. Por ejemplo, $R^2 = 0{,}5$ sólo dice que el 50% de la varianza de las observaciones queda explicado por el modelo lineal.

La relación entre R^2 y r ayuda a comprender lo expuesto en el análisis de la correlación: que un valor de $r^2 = 0{,}5$ indica una correlación débil. Este valor representará un $R^2 = 0{,}25$; es decir, el modelo de regresión sólo explica un 25% de la variabilidad total de las observaciones.

El signo de r da información de si la relación es positiva o negativa. Así pues, con el valor de r siempre se puede calcular el valor de R^2, pero al revés quedará indeterminado el valor del signo a menos que conozcamos la pendiente de la recta. Por ejemplo, dado un $R^2 = 0{,}81$, si se sabe que la pendiente de la recta de regresión es negativa, entonces se puede afirmar que el coeficiente de correlación r será igual a 0,9.

Predicción

La predicción constituye una de las aplicaciones más interesantes de la técnica de regresión. La predicción consiste en determinar a partir del modelo estimado el valor que toma la variable endógena para un valor determinado de la exógena. La fiabilidad de esta predicción será tanto mayor, en principio, cuanto mejor sea el ajuste (es decir, cuanto mayor sea R^2), en el supuesto de que exista relación causal entre la variable endógena y la variable exógena.

Nota

Variable endógena es la variable dependiente. Es la variable que se predice o se explica. Se representa por *Y*.

Variable exógena es la variable independiente. Es la variable que sirve para predecir o explicar. Se representa por *X*.

Ejemplo 1. Estudio de los servicios ofrecidos por un centro de documentación.

Una vez obtenida la ecuación estimada de regresión $\hat{y} = 1{,}4 + 0{,}5x$ del ejemplo anterior, interpretamos los resultados:

En este caso la ordenada en el origen ($\hat{\beta}_0 = 1{,}4$) si puede tener interpretación con sentido, ya que correspondería a la estimación de la puntuación obtenida para la atención al usuario cuando la puntuación del funcionamiento global es cero. La pendiente ($\hat{\beta}_1 = 0{,}5$) es positiva, lo que indica que el aumento en una unidad de la valoración del funcionamiento global del centro está asociado con un aumento de 0,5 unidades en la puntuación de atención al usuario.

Si quisiéramos predecir la valoración de la atención para una persona que ha valorado 7 el funcionamiento global, el resultado sería:

$$\hat{y} = 1{,}4 + 0{,}5 \cdot 7 = 4{,}9$$

En el ejemplo hemos obtenido la ecuación de regresión y debemos analizar la bondad de dicho ajuste que daría respuesta a la siguiente pregunta: ¿se ajustan bien los datos a esta ecuación de regresión?

Calcularemos el coeficiente de determinación que es una medida de la corrección del ajuste. Para ello tenemos que descomponer la variabilidad total de las observaciones de la forma:

$$SST = SSR + SSE$$

Utilizando los valores de la tabla 2 (cálculo de las sumas de cuadrados para la ecuación estimada de regresión con mínimos cuadrados), calculamos *SST* = suma de cuadrados total, es la suma de la última columna de la tabla 2.

$$SST = \sum_{i=1}^{5} (y_i - \bar{y})^2 = 13,2$$

En la tabla 3 vemos los cálculos necesarios para determinar la *SSE* = suma de cuadrados debida al error

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} e_i^2 = 3,2$$

Tabla 3. Cálculo de las sumas de cuadrados debidas al error SCE

Funcionamiento (X)	**Atención (Y)**	$\hat{y} = 1,4 + 0,5x_i$	$e = y_i - \hat{y}_i$	$(y_i - \hat{y}_i)^2$
2	2	2,4	–0,4	0,16
4	4	3,4	0,6	0,36
6	5	4,4	0,6	0,36
8	4	5,4	–1,4	1,96
10	7	6,4	0,6	0,36

$$SSE = \sum_{i=1}^{5} (y_i - \hat{y}_i)^2 = 3,2$$

La *SSR* = suma de cuadrados debida a la regresión se puede calcular con facilidad usando esta expresión:

$$SSR = \sum_{i=1}^{5} (\hat{y}_i - \bar{y})^2$$

o bien si se conocen *SST* y *SSE* se puede obtener fácilmente.

$$SSR = SST - SSE = 13,2 - 3,2 = 10$$

El valor del coeficiente de determinación será:

$$R^2 = \frac{SSR}{SST} = \frac{10}{13,2} = 0,7576$$

Si lo expresamos en porcentaje, R^2= 75,76 %. Podemos concluir que el 75,76% de la variación de la puntuación en la atención al usuario se puede explicar con la relación lineal entre las valoraciones del funcionamiento global del centro y la atención al usuario. El ajuste al modelo lineal es bueno. Se considera un buen ajuste cuando R^2 es mayor o igual que 0,5.

El coeficiente de correlación lineal "*r*" será $\sqrt{0{,}75760} = |0{,}87|$, resultado acorde con la estimación obtenida usando la covarianza.

Solución de problemas de regresión lineal simple con programas informáticos

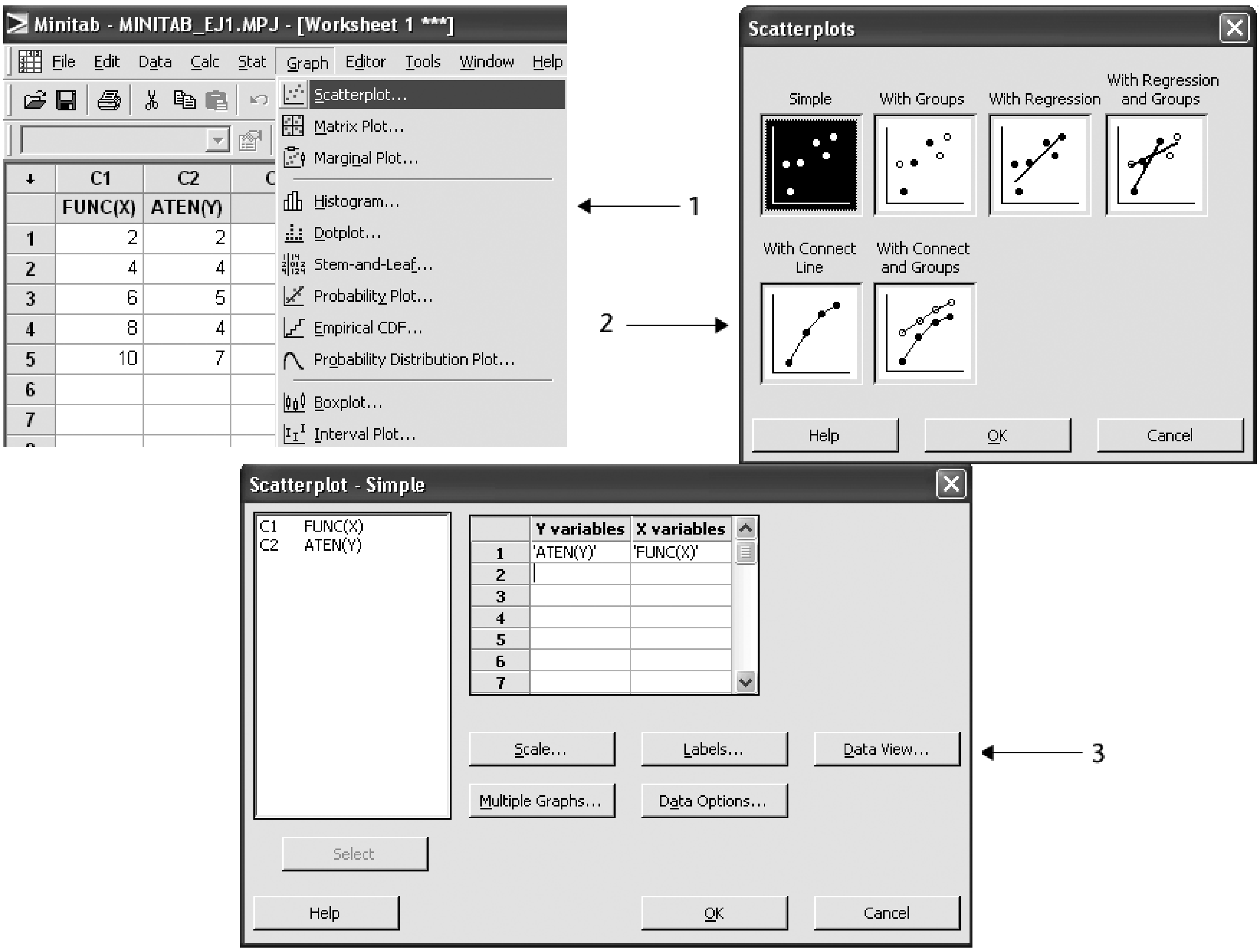
Para resolver el ejercicio empleamos el programa Minitab.

Insertamos los datos del ejemplo 1: "Estudio de los servicios ofrecidos por un centro de documentación". A la variable independiente (*Y*) la llamamos **ATEN** (de atención al usuario) y a la variable dependiente (*X*) la llamamos **FUNC** (de funcionamiento global) para facilitar la interpretación de los resultados. Insertamos los datos **FUNC** en la columna C1 y los datos de **ATEN** en la columna C2, con encabezados para obtener el diagrama de dispersión.

Pasos a seguir

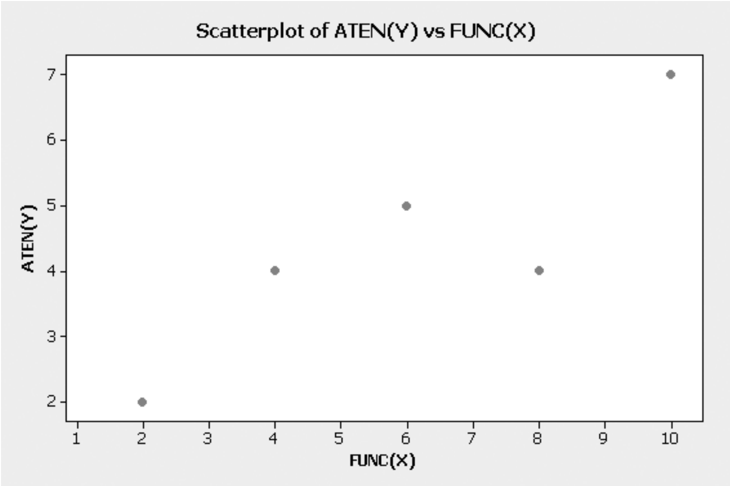
Para crear el gráfico una vez introducidos los datos en el programa ***(1)***, se sigue la ruta ***Graph*** > ***Scatterplot*** > ***Simple*** ***(2)*** y se rellenan los campos en la ventana correspondiente seleccionando las variables ***(3)***. Seleccionad ***OK*** para obtener el diagrama de dispersión.

Figura 5. Pasos a seguir para obtener el diagrama de dispersión



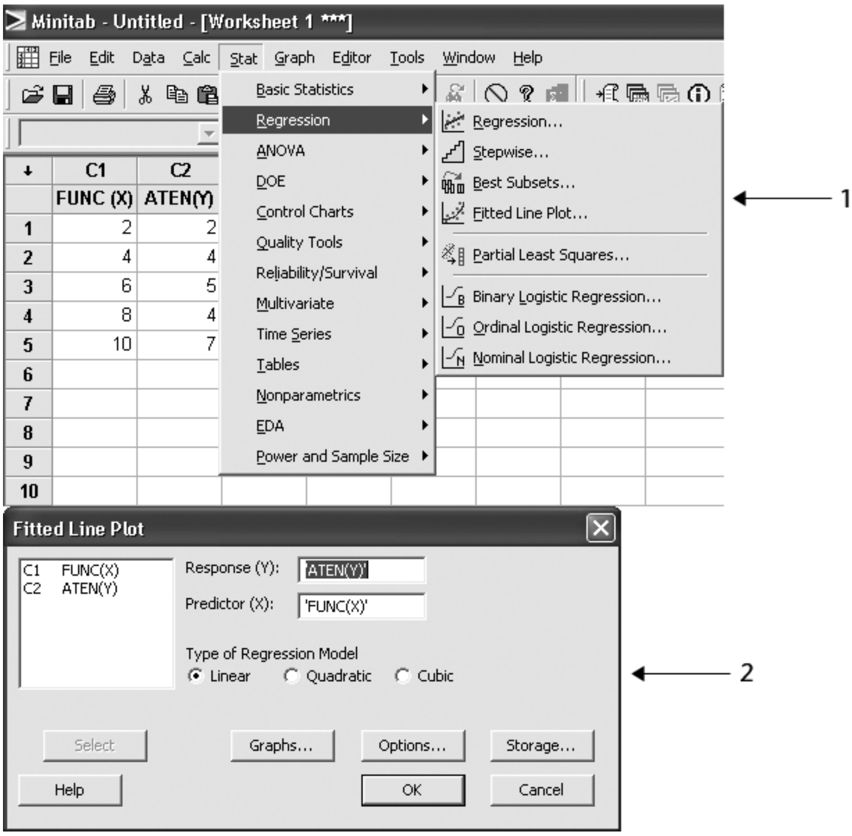
Obtuvimos el diagrama de la figura 6.

Figura 6. Diagrama de dispersión. Minitab



La figura 7 muestra los pasos a seguir para representar la recta de de regresión de mínimos cuadrados:

Figura 7. Pasos a seguir para representar la recta de regresión de mínimos cuadrados



Pasos a seguir

Usamos la opción ***Stat***, se sigue la ruta ***Regression*** > ***Regression*** > ***Fitted Line Plot (1)*** y se rellenan los campos en la ventana correspondiente ***(2)***. Seleccionad ***OK*** para obtener el gráfico.

Obtuvimos los resultados que aparecen en la figura 8.

A continuación interpretaremos los resultados:

La figura 8 muestra la gráfica de la ecuación de regresión sobre el diagrama de dispersión. La pendiente de la ecuación de regresión ($\hat{\beta}_1 = 0{,}50$) es positiva, lo

que implica que al aumentar las valoraciones del funcionamiento global, las puntuaciones de atención al usuario también aumentan.

Figura 8. Gráfica de la ecuación de regresión de mínimos cuadrados

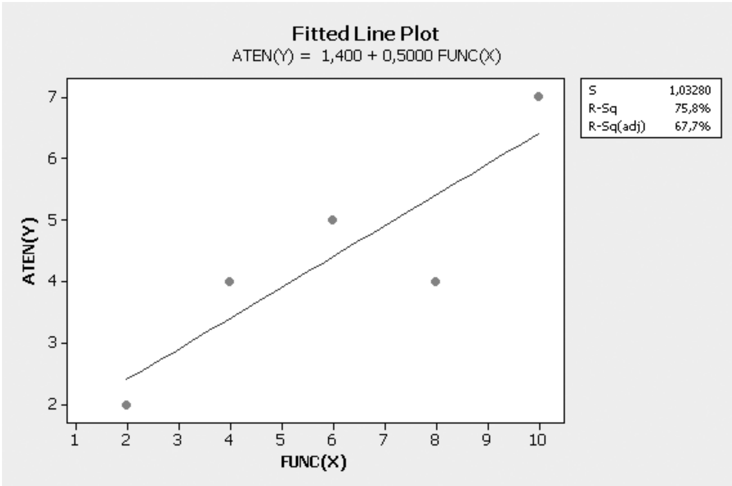
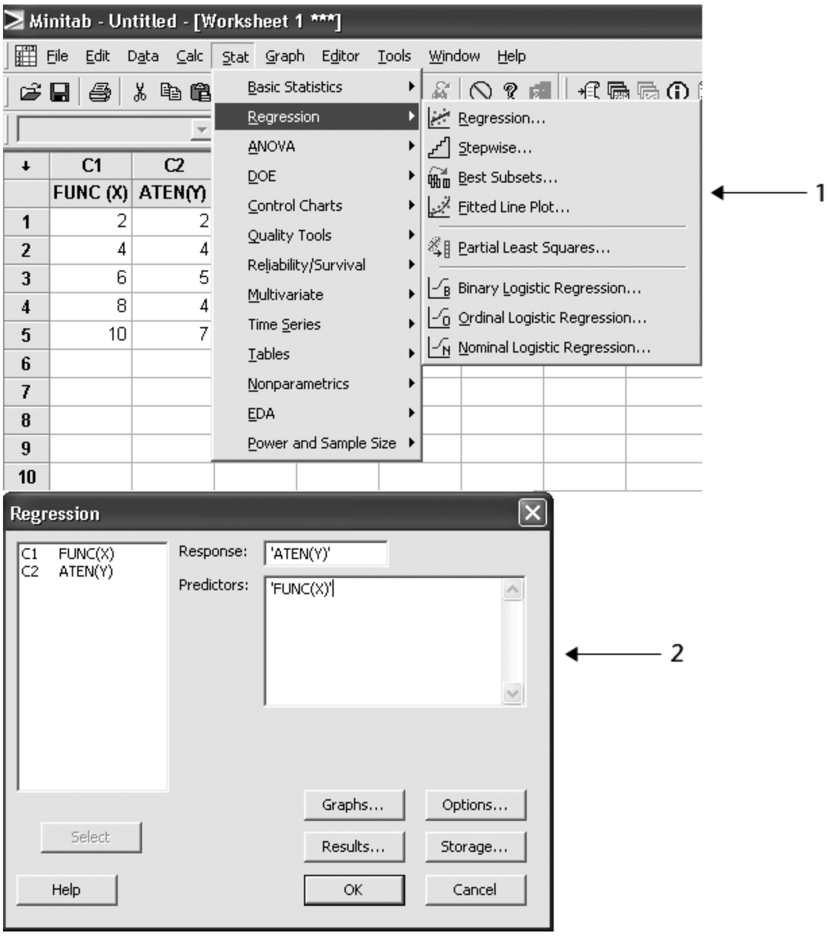


Figura 9. Pasos a seguir para realizar el análisis de regresión



Pasos a seguir

Se sigue la ruta ***Stat*** > ***Regression*** > ***Regression (1)*** y se rellenan los campos en la ventana correspondiente ***(2)***. Seleccionad ***OK*** para obtener el análisis de regresión.

En el cuadro de diálogo de Minitab puede obtenerse más información sobre resultados seleccionando las opciones deseadas. Por ejemplo, con este cuadro de diálogo se pueden obtener los residuos, los residuales estandarizados, los puntos de alta influencia y la matriz de correlación (estos resultados los comentaremos más adelante).

Obtenemos los resultados que aparecen en la figura 10.

Figura 10. Resultados del análisis de regresión. Minitab

```
Regression Analysis: ATEN(Y) versus FUNC(X)

The regression equation is
  ATEN(Y) = 1,40 + 0,500 FUNC(X) <------------  Ecuación estimada de
                                                regresión

Predictor  Coef  SE Coef    T      P    --
Constant   1,400  1,083    1,29  0,287   }-   Prueba T
FUNC(X)    0,5000  0,1633  3,06  0,055  --
S = 1,03280   R-Sq = 75,8%   R-Sq(adj) = 67,7%

Analysis of Variance

Source          DF     SS      MS     F     P   --
Regression       1   10,000  10,000  9,38  0,055  }-  Tabla
ANOVA                                             --
Residual Error   3   3,200   1,067
Total            4  13,200

  Predicted Values for New Observations

New
Obs    Fit   SE Fit    95% CI          95% PI
  1    4.900  0.490    (3.341; 6.459)   (1.262; 8.538) <----Estimados del
                                                            intervalo

Values of Predictors for New Observations

New
Obs  FUNC(X)
  1     7.00
```

- Interpretación de las estadísticas de regresión:

  Minitab imprime la ecuación de regresión en la forma:

  $$\text{ATEN}(Y) = 1{,}40 + 0{,}500\ \text{FUNC}(X).$$

  Se imprime una tabla que muestra los valores de los coeficientes a y b. El coeficiente *Constant* (ordenada en el origen) es 1,4, y la pendiente con base en la variable FUNC es 0,50. *SE Coef* son las desviaciones estándar de cada coeficiente. Los valores de las columnas T y P los analizaremos más adelante al estudiar la inferencia en la regresión.

  El programa imprime el error estándar del valor estimado, $S = 1{,}03280$ mide el tamaño de una desviación típica de un valor observado (x,y) a partir de la recta de regresión. También proporciona la información sobre la bondad de ajuste. Observad que *R-Sq* = 75,8% ($R^2 = 0{,}758$) es el coeficiente de determinación expresado en porcentaje. Como hemos comentado en la solución manual del ejercicio, un valor del 75,8% significa que el 75,8% de la variación en la puntuación de atención al usuario puede explicarse por medio de la valoración obtenida en el funcionamiento global del centro. Se supone que el 24,2 % restante de la variación se debe a la variabilidad aleatoria. El resultado *R-Sq*(adj) = 67,7% (R^2 ajustado) es un valor corregido de

acuerdo con la cantidad de variables independientes. Se tiene en cuenta al realizar una regresión con varias variables independientes y se estudiará más adelante al tratar la regresión múltiple.

- Interpretación del análisis de la varianza:

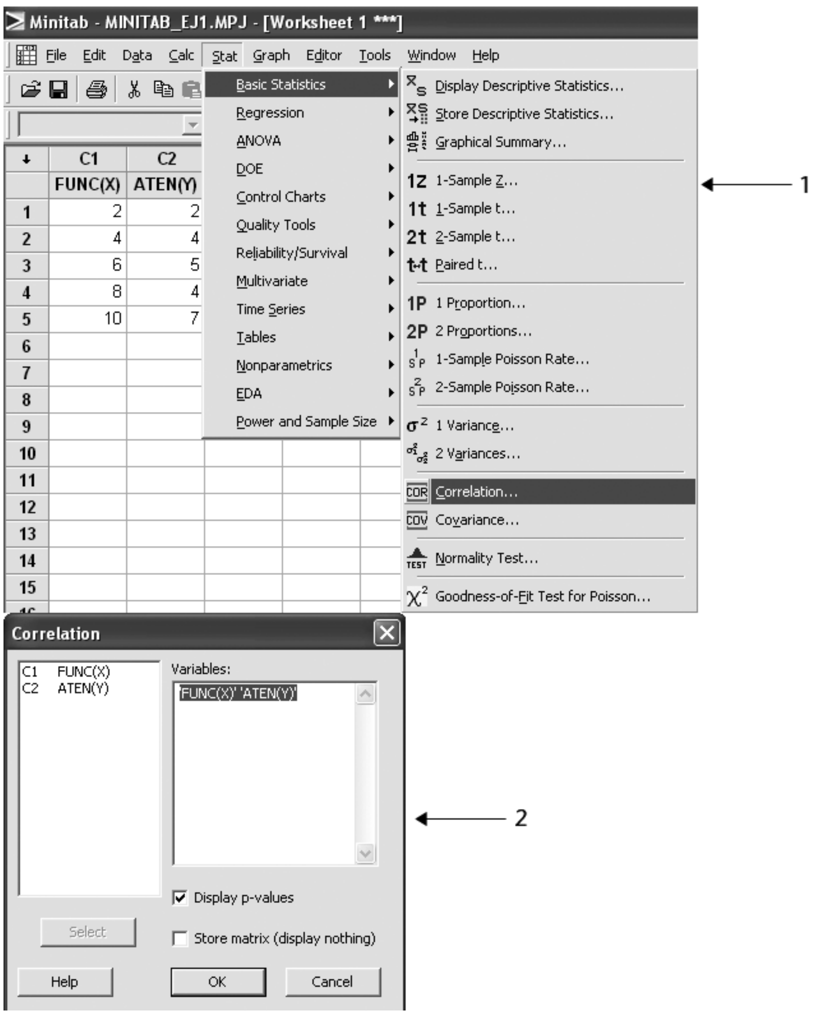
  La salida de Minitab analiza la variabilidad de las puntuaciones de atención al usuario. La variabilidad, como hemos explicado anteriormente, se divide en dos partes: *SST* = *SSR* + *SSE*.

  SS Regresión (*SSR*) es la variabilidad debida a la regresión, ***SS Error*** (*SSE*) es la variabilidad debida al error o variabilidad aleatoria, ***SS Total*** (*SST*) es la variabilidad total. El resto de la información se irá viendo mas adelante al tratar la regresión lineal múltiple.

- Interpretación del valor estimado de predicción y del intervalo de confianza de 95% (95% C.I.) y el estimado del intervalo de predicción (95% P.I.) de la atención al usuario para el valor 7 de funcionamiento global. El valor estimado para Atención al usuario es 4,9.

  A continuación calcularemos el coeficiente de correlación lineal como se indica en la figura 11.

Figura 11. Pasos a seguir para calcular el coeficiente de correlación



Pasos a seguir

Para crear el gráfico se sigue la ruta ***Stat*** > ***Basic Statistics*** > ***Correlation (1)*** y se rellenan los campos en la ventana correspondiente ***(2)***. Seleccione ***OK*** para obtener el coeficiente de correlación lineal.

Obtuvimos los resultados que aparecen en la figura 12.

Figura 12. Resultados del análisis de correlación

Correlations: FUNC(X); ATEN(Y)

```
Pearson correlation of FUNC(X) and ATEN(Y) = 0,870
P-Value = 0,055
```

• Interpretación del análisis de correlación:

Como *r* =0,870, podemos decir que existe correlación lineal positiva entre las valoraciones obtenidas de atención al usuario y las valoraciones del funcionamiento global del centro. El funcionamiento está asociado positivamente con la atención al usuario.

Obsérvese que $R^2 = 0,758$, por lo que $\sqrt{R^2} = \sqrt{0,758} = 0,87 = r$

Para resolver el ejemplo 1. "Estudio de los servicios ofrecidos por un centro de documentación" se emplea **Microsoft Excel**.

La figura 13 muestra el correspondiente *output* que ofrece **Microsoft Excel**.

Se observa que las estadísticas de regresión coinciden con las obtenidas con Minitab.

Atención

Para poder hacer la regresión con **MS Excel** es necesario instalar previamente un complemento llamado "Análisis de datos". Para instalar las herramientas de análisis de datos, haced clic en ***Herramientas*** > ***Complementos,*** y en el cuadro de diálogo activar: ***Herramientas para análisis.***

Figura 13. Resultados del análisis de regresión del ejemplo 1. "Estudio de los servicios ofrecidos por un centro de documentación". Excel

	A	B	C	D	E	F	G	H	I
1	Resumen								
2									
3	*Estadísticas de la regresión*								
4	Coeficiente de correlación múltiple	0,87038828							
5	Coeficiente de determinación R^2	0,757575758							
6	R^2 ajustado	0,676767677							
7	Error típico	1,032795559							
8	Observaciones	5							
9									
10	ANÁLISIS DE VARIANZA								
11		*Grados de libertad*	*Suma de cuadrados*	*Promedio de los cuadrados*	*F*	*Valor crítico de F*			
12	Regresión	1	10	10	9,375	0,054912524			
13	Residuos	3	3,2	1,066666667					
14	Total	4	13,2						
15									
16		*Coeficientes*	*Error típico*	*Estadístico t*	*Probabilidad*	*Inferior 95%*	*Superior 95%*	*Inferior 95,0%*	*Superior 95,0%*
17	Intercepción	1,4	1,083205121	1,292460655	0,28674468	-2,047242134	4,847242134	-2,047242134	4,847242134
18	Funcionamiento (X)	0,5	0,163299316	3,061862178	0,05491252	-0,019691305	1,019691305	-0,019691305	1,019691305
19									
20									
21									
22	Análisis de los residuales				Resultados de datos de probabilidad				
23									
24	*Observación*	*Pronóstico Atención (Y)*	*Residuos*		*Percentil*	*Atención (Y)*			
25	1	2,4	-0,4		10	2			
26	2	3,4	0,6		30	4			
27	3	4,4	0,6		50	4			
28	4	5,4	-1,4		70	5			
29	5	6,4	0,6		90	7			
30									

Diagnóstico de la regresión

Al igual que en cualquier procedimiento estadístico, cuando se efectúa una regresión en un conjunto de datos se hacen algunas suposiciones importantes, y en este caso son cuatro:

1) El modelo de línea recta es correcto.

2) Los errores o residuos siguen una distribución aproximadamente normal de media cero.
3) Los errores o residuos tienen una varianza constante σ^2.
4) Los errores o residuos son independientes.

Siempre que usen regresiones para ajustar una recta a los datos, deben considerarse estas suposiciones. Comprobar que los datos cumplen estas suposiciones supone pasar por una serie de pruebas llamadas **diagnosis** que se describen a continuación.

Prueba de suposición de línea recta.

Para comprobar si es correcto el modelo de línea recta se usa el gráfico de dispersión con el ajuste a la recta de mínimos cuadrados (ejemplo 1, figura 14).

Figura 14. Gráfica de la ecuación de regresión del ejemplo 1



Análisis de residuos

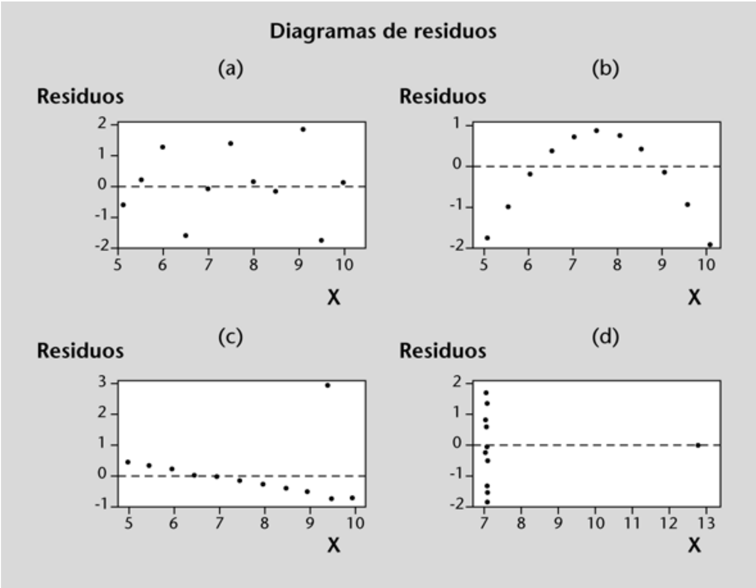
Una vez hecho el ajuste de un modelo de regresión lineal a los datos muestrales, hay que efectuar el análisis de los residuos o errores. Este análisis, que a continuación comentaremos de forma breve e intuitiva, nos servirá para hacer un diagnóstico del modelo de regresión.

Otra forma de ver si los datos se ajustan a una recta es realizando un gráfico de los residuos ($e_i = y_i - \hat{y}_i$) en función de la variable predictora (X). En el eje horizontal se representa el valor de la variable independiente (X) y en el vertical los valores de los residuos (e_i).

Podemos calcular los residuos manualmente según habíamos indicado en la tabla 3.

En la figura 15 presentamos 4 ejemplos de gráficos de residuos o errores.

Figura 15. Diagrama de residuos



Podemos observar que de los cuatro, sólo el primero no presenta ningún tipo de estructura, los residuos se distribuyen aleatoriamente, de manera que sólo tendría sentido la regresión hecha sobre la muestra (a). Si los puntos se orientasen en forma de "U" (o "U" invertida), habría problemas con este supuesto, como es el caso de la muestra (b). Los residuos del diagrama (c) y (d) no se distribuyen aleatoriamente, por lo que no se cumple el supuesto de linealidad.

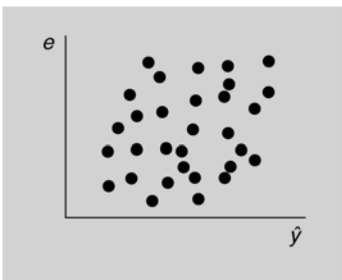
En el mismo gráfico también podemos observar si los residuos tienen varianza constante (supuesto 3). Si la varianza de los errores es constante para todos los valores de X, la gráfica de residuales debe mostrar un patrón similar a una banda horizontal de los puntos, como en (a). Si forman una flecha (en un extremo se agrupan mucho más que en el otro), caso (d), entonces este supuesto falla. También es conveniente estar atentos ante la posible existencia de valores atípicos o valores extremos (*outliers*), pues éstos podrían afectar.

Valor atípico

Por *valor atípico* entendemos un valor muy diferente de los otros y que muy posiblemente es erróneo.

También podemos usar un gráfico de residuos en función del valor estimado o predicho $\hat{y}$. Esto lo representaremos gráficamente mediante un diagrama de dispersión de los puntos $(\hat{y}_i, e_i)$, es decir, sobre el eje de las abscisas representamos el valor estimado $\hat{y}$, y sobre el eje de ordenadas, el valor correspondiente del residuo, es decir, $e_i = y_i - \hat{y}_i$.

Figura 16. Gráfico de residuos en función de valor estimado o predicho $\hat{y}$



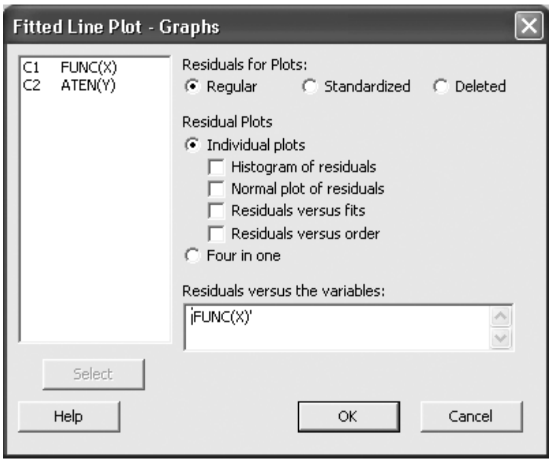
Si el modelo lineal obtenido se ajusta bien a los datos muestrales, entonces la nube de puntos $\left(\hat{y}_i, e_i\right)$ no debe mostrar ningún tipo de estructura. Para la regresión lineal simple, la gráfica de residuos en función de X y los de residuos en función de $\hat{y}$ dan la misma información. Para la regresión múltiple, la gráfica de residuos en función de $\hat{y}$ se usa con más frecuencia porque se maneja más de una variable independiente.

Para comprobar el segundo supuesto de que los errores o residuos siguen una distribución aproximadamente normal usaremos la gráfica de probabilidad normal.

Consideramos de nuevo el ejemplo 1. "Estudio de los servicios ofrecidos por un centro de documentación" y realizamos la diagnosis con Minitab a fin de comprobar si se cumplen las condiciones del modelo.

En la figura 17 se indican los pasos a seguir para crear un gráfico de los residuos en función de la variable de predicción con Minitab:

Figura 17. Pasos a seguir para crear un gráfico de los residuos en función de la predicción

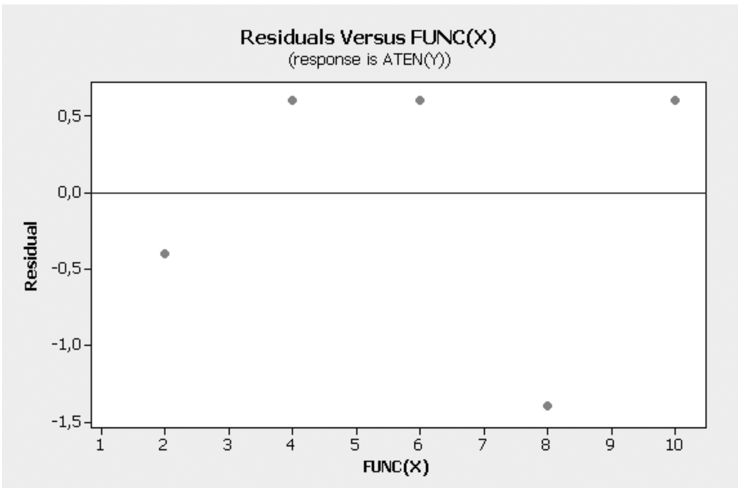


Pasos a seguir

Se sigue la ruta ***Stat*** > ***Regression*** > ***Fitted Line Plot*** > ***Linear*** > ***Graph*** y se rellenan los campos correspondientes. Seleccione ***OK*** para obtener el gráfico de residuos.

Obtenemos la gráfica que aparece en la figura 18.

Figura 18. Gráfica de los residuos en función de la variable independiente

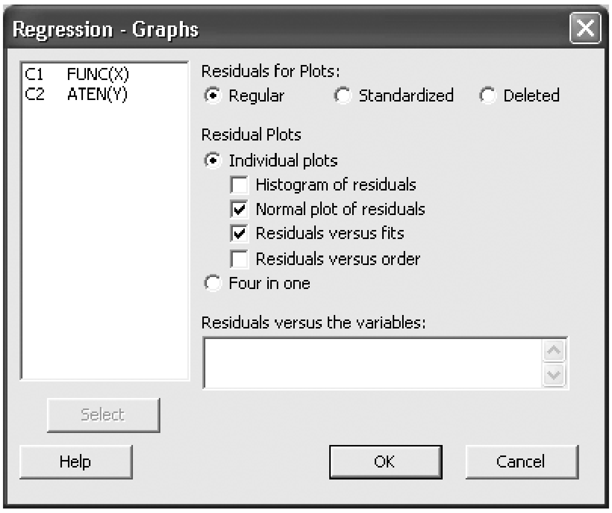


Los valores residuales se distribuyen aleatoriamente y no presenta ningún tipo de estructura, por consiguiente concluimos que la gráfica de los residuos no muestra evidencia de incumplir el supuesto de linealidad y podemos por ahora concluir que el modelo lineal simple es válido para el ejemplo "Estudio de los servicios ofrecidos por un centro de documentación".

En el mismo gráfico podemos observar que los residuos tienen varianza constante ya que parecen estar en la banda horizontal.

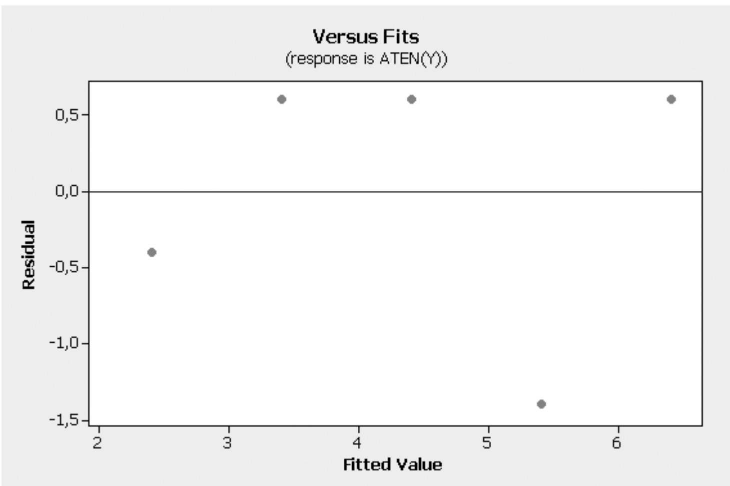
A fin de comprobar si se cumplen el resto de las condiciones del modelo, seleccionamos la opción ***Graphs*** y completamos los campos según se indica en la figura 19:

Figura 19. Pasos a seguir para crear un gráfico de los residuos en función de los valores estimados (fits)



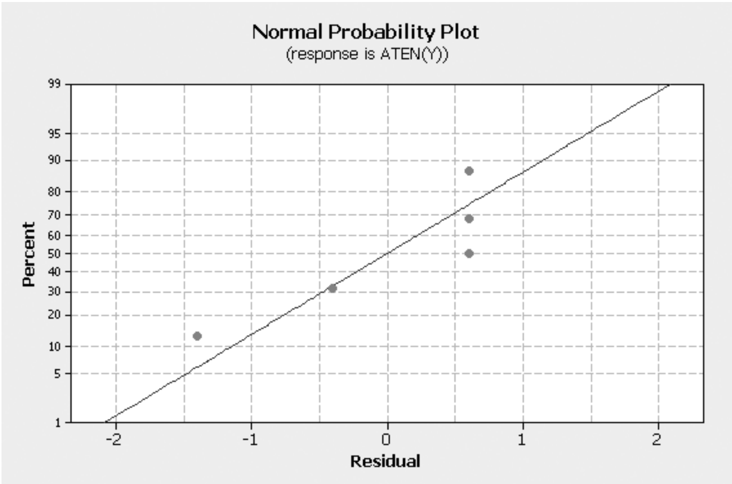
La figura 20 presenta el gráfico de los valores residuales frente a los valores estimados y el significado es análogo al de la figura 18. Los residuos se distribuyen aleatoriamente, no presenta ningún tipo de estructura, y podemos concluir que es válido el modelo lineal simple.

Figura 20. Gráfica de los residuos en función de los valores estimados



En la gráfica de la figura 21 podemos comprobar que los residuos siguen una distribución aproximadamente normal, ya que los puntos se acercan bastante a una recta (esta hipótesis sólo plantearía dificultades si estos puntos se alejasen de la forma lineal):

Figura 21. Gráfica de probabilidad normal



Inferencia en la regresión: constrastes de hipótesis e intervalos de confianza

Al hacer un análisis de regresión se comienza proponiendo una hipótesis acerca del modelo adecuado de la relación entre las variables dependiente e independiente. Para el caso de regresión lineal simple, el modelo de regresión supuesto es:

$$y = \beta_0 + \beta_1 x_i + \varepsilon_i$$

A continuación aplicamos el método de mínimos cuadrados para determinar los valores de los estimadores $\hat{\beta}_0$ y $\hat{\beta}_1$ de los parámetros del modelo. La ecuación estimada de regresión que resulta es:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

Ya hemos visto que el valor del coeficiente de determinación (R^2) es una medida de bondad de ajuste de esta ecuación. Sin embargo, aun con un valor grande de R^2 no se debería usar la ecuación de regresión sin antes efectuar un análisis de la adecuación del modelo propuesto. Para ello se debe determinar el significado (o importancia estadística) de la relación. Las pruebas de significación en el análisis de regresión se basan en los siguientes supuestos acerca del término del error ε:

1) El término del error ε es una variable aleatoria con distribución normal con media, o valor esperado, igual a cero.

2) La varianza del error, representada por σ^2, es igual para todos los valores de x.

3) Los valores de los errores son independientes.

Base para la inferencia sobre la pendiente de la regresión poblacional

Sea β_1 la pendiente del modelo de regresión y $\hat{\beta}_1$ su estimación por mínimos cuadrados (basada en observaciones muestrales). Si se cumplen los supuestos acerca del término del error expuestos anteriormente, la pendiente del modelo de regresión β_1 se distribuye como una *t* de Student con (*n* – 2) grados de libertad.

$$t = \frac{\hat{\beta}_1 - \beta_1}{S_{\hat{\beta}_1}}$$

Para obtener el estadístico de contraste, calcularemos:

$S_{\hat{\beta}_1}$ es la desviación estándar estimada de β_1,

$$s_{\hat{\beta}_1} = \frac{s}{\sqrt{\sum_{i}^{n} (x_i - \bar{x})^2}}$$

s es el error estándar de los estimados. Para calcularlo, se divide la suma de las desviaciones al cuadrado por *n* – 2, que son los grados de libertad.

$$s = \frac{1}{n-2} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

En el análisis de regresión aplicado, primero se desea conocer si existe una relación entre las variables *X* e *Y*. En el modelo se ve que si β_1 es cero, entonces no existe relación lineal: *Y* no aumentaría o disminuiría cuando aumenta *X*. Para averiguar si existe una relación lineal, se puede contrastar la hipótesis

$$H_0:\ \beta_1 = 0$$

frente a

$$H_1:\ \beta_1 \neq 0$$

Se puede contrastar esta hipótesis utilizando el estadístico *t* de Student

$$t = \frac{\hat{\beta}_1 - \beta_1}{S_{\hat{\beta}_1}} = \frac{\hat{\beta}_1 - 0}{S_{\hat{\beta}_1}} = \frac{\hat{\beta}_1}{S_{\hat{\beta}_1}}\ ,$$

que se distribuye como una t de Student con $n - 2$ grados de libertad. La mayoría de los programas que se emplean para estimar regresiones la desviación estándar de los coeficientes y el estadístico t de Student para $\beta_1 = 0$. Las figuras 10 y 13 muestran respectivamente las salidas de Minitab y Excel correspondientes al ejemplo del estudio de los servicios ofrecidos por un centro de documentación.

En el caso del modelo de ejemplo, el coeficiente de la pendiente es $\hat{\beta}_1 = 0,50$ con una desviación estándar $S_{\hat{\beta}_1} = 0,1633$. Para saber si existe relación entre la atención al usuario, Y, y el funcionamiento global, X, se puede contrastar la hipótesis $H_0 : \beta_1 = 0$ frente a $H_1 : \beta_1 \neq 0$. Este resultado se obtiene en el caso de un contraste de dos colas con un nivel de significación $\alpha = 0,05$ y 3 grados de libertad.

El estadístico t calculado es:

$$t = \frac{0,50 - 0}{0,1633} = 3,06$$

Recordad

El p-valor es la probabilidad de que una variable aleatoria supere el valor observado para el estadístico de contraste.

- Si p-valor $< \alpha$, se rechaza H_0.
- Si p-valor $\geq \alpha$, no se rechaza H_0.

El estadístico t resultante, $t = 3,06$, mostrado en la salida de regresión de la figura 22, es la prueba definitiva para rechazar o aceptar la hipótesis nula. En este caso el p-valor es 0,055; como p-valor $> 0,05$ (no podemos rechazar la H_0: $\beta_1 = 0$ al nivel de significación de $\alpha = 0,05$), se acepta que $\hat{\beta}_1 = 0$. Por lo tanto, no se puede afirmar que exista una relación lineal entre las valoraciones del funcionamiento global y la atención al usuario a un nivel de confianza del 95% (nivel de significación del 0,05).

Figura 22. Resumen de la figura 10. Resultados del análisis de regresión. Minitab

Regression Analysis: ATEN(Y) versus FUNC(X)

```
The regression equation is
ATEN(Y) = 1,40 + 0,500 FUNC(X)
Predictor     Coef   SE Coef      T       P
Constant     1,400     1,083   1,29   0,287
FUNC(X)     0,5000    0,1633   3,06   0,055

S = 1,03280   R-Sq = 75,8%   R-Sq(adj) = 67,7%
```

Si el nivel de significación se hubiera fijado del 10% ($\alpha = 0,10$), se podría rechazar H_0, ya que el p-valor $< 0,10$, los resultados indicarían que $\beta_1 \neq 0$ y en este caso se podría decir que a un nivel de confianza del 90% existe relación lineal entre ambas variables.

Intervalo de confianza para la pendiente

Se puede obtener intervalos de confianza para la pendiente β_1 del modelo de regresión utilizando los estimadores de los coeficientes y de las varianzas que se han desarrollado y el razonamiento utilizado en el módulo 2.

Si los errores de la regresión ε_i siguen una distribución normal y se cumplen los supuestos de la regresión, se obtiene un intervalo de confianza al (1 – α)% de la pendiente del modelo de regresión simple β_1 de la siguiente forma:

$$\hat{\beta}_1 - t_{n-2,\alpha/2} s_{\hat{\beta}_1} < \beta_1 < \hat{\beta}_1 + t_{n-2,\alpha/2} s_{\hat{\beta}_1}$$

donde $t_{n-2,\alpha/2}$ es el número para el que

$$P(t_{n-2} > t_{n-2,\alpha/2}) = \alpha/2$$

el estadístico t_{n-2} sigue una distribución *t* de Student con (n – 2) grados de libertad.

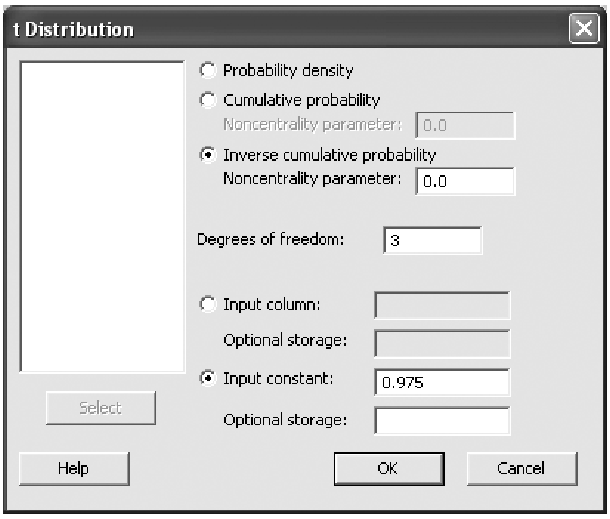
En la salida del análisis de regresión de la atención al usuario respecto al funcionamiento global del centro de documentación de la figura 22, se observa que

$$n = 5 \quad \hat{\beta}_1 = 0{,}50 \quad S_{\hat{\beta}_1} = 0{,}1633$$

Para obtener el intervalo de confianza al 95% de β_1, $(1 - \alpha) = 0{,}95$ y $n - 2 = 3$ grados de libertad, es necesario calcular el valor crítico de la *t*-Student. En este caso con $n - 2 = 5 - 2 = 3$ grados de libertad y $\alpha/2 = 0{,}05/2 = 0{,}025$. Se puede obtener utilizando las tablas de la distribución *t* de Student o con el ordenador.

Si se utiliza Minitab, los pasos a seguir se muestran en la figura 23.

Figura 23. Pasos a seguir para calcular el valor crítico *t*



Pasos a seguir

Se sigue la ruta ***Calc*** > ***Probability Distributions*** > ***t*** y se rellenan los campos en la ventana correspondiente. Seleccionad ***OK*** para obtener el *output* de la figura 24.

Figura 24. Resultados de cálculo del valor crítico *t*. Minitab

Inverse Cumulative Distribution Function

```
Student's t distribution with 3 DF

P( X <= x )        x
      0.975  3.18245
```

el valor de $t_{n-2,\alpha/2} = t_{3;0,025} = 3,18$

Por lo tanto, el intervalo de confianza al 95% será

$$0,50 - (0,1633)\,(3,18) < \beta_1 < 0,50 + (0,1633)\,(3,18)$$

O sea

$$-0,019 < \beta_1 < 1,0193$$

Por tanto, el intervalo de confianza buscado es: $0,50 \pm 3,18245 \cdot 0,1633$, *i. e.*, se puede afirmar con una probabilidad del 95% que β_1 se encuentra en el intervalo de extremos –0,0197 y 1,0197.

En la tabla 4 se presentase el intervalo de confianza calculado con Excel. El resumen muestra en las ultimas columnas los valores estimados de intervalo de confianza del 95% para los parámetros de regresión β_0 y β_1, también las desviaciones estándar estimadas (columna *Error típico*), el valor estadístico *t* (columna *Estadístico t*) y los *p*-valores (columna *Probabilidad*).

Tabla 4. Resumen de la figura 13 (Resultados del análisis de regresión. Excel)

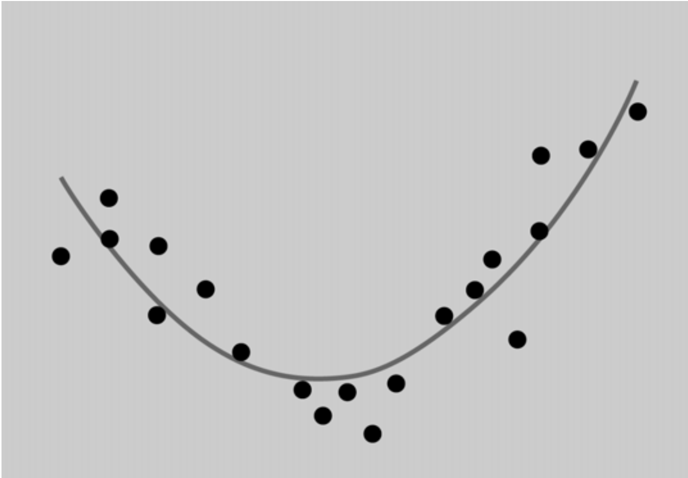
	Coeficientes	Error típico	Estadístico *t*	Probabilidad	Inferior 95%	Superior 95%
Intercepción	1,4	1,08320512	1,29246066	0,286745	–2,047242	4,847242134
Funcionamiento (*X*)	0,5	0,16329932	3,06186218	0,054913	–0,019691	1,019691305

3.2. Modelos de regresión simple no lineales: modelo cuadrático y cúbico

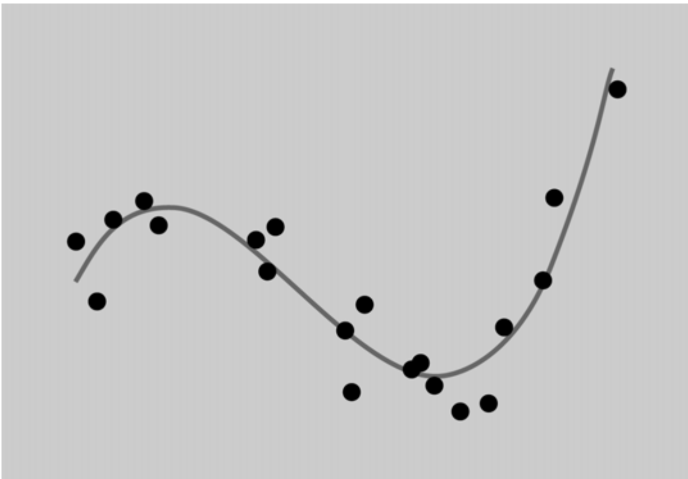
Existen algunas relaciones que no son estrictamente lineales, y se pueden desarrollar métodos con el fin de poder utilizar los métodos de regresión para estimar los coeficientes del modelo.

Aparte de los modelos de regresión lineales, se pueden establecer otros que no son lineales, entre los cuales destacamos: el modelo cuadrático y el cúbico, que son modelos curvilíneos. Cada modelo corresponde con el grado de la ecuación, siendo *Y* la respuesta y *X* la variable predictora, β_0 la ordenada en el origen, y β_1, β_2, y β_3 los coeficientes. Es importante escoger el modelo apropiado cuando se modelizan datos usando regresión y análisis de tendencia.

Modelo cuadrático: $Y = \beta_0 + \beta_1 X + \beta_2 X^2$



Modelo cúbico: $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3$



Para determinar qué modelo utilizar, se representan previamente los datos (diagrama de dispersión) y se calcula el coeficiente de correlación lineal de Pearson. Conviene recordar que dicho coeficiente "r" mide el grado de asociación que existe entre las variables X e Y cuando se ajusta a su nube de puntos una *línea recta,* pero no mide el grado de ajuste de una curva a la nube de puntos. Podría darse el caso de que la relación entre las variables fuera grande, sólo que distribuida a lo largo de una curva, en cuyo caso, al ajustar a una recta se obtendría un coeficiente de correlación lineal "r" y un coeficiente de determinación "R^2" bajo. Calcularíamos el ajuste simultáneo a los modelos no lineales (cuadrático y cúbico) y se calcularían los coeficientes de determinación para ambos modelos para determinar la bondad del ajuste. El mejor modelo será el que presente el valor más elevado de R^2.

Los métodos de inferencia para los modelos no lineales transformados son los mismos que se han desarrollado para los modelos lineales. Así, si se tiene un modelo cuadrático, el efecto de una variable X está indicado por los coeficientes tanto de los términos lineales como de los términos cuadráticos.

Ejemplo. Número de visitantes a un museo (estimación de un modelo cuadrático utilizando Minitab)

Se desea estudiar la variación entre el número de visitantes a un museo en función del número de obras visitadas. La tabla 5 muestra el número de visitantes y el número de obras visitadas. Se han seleccionado aleatoriamente los datos correspondientes a 6 días.

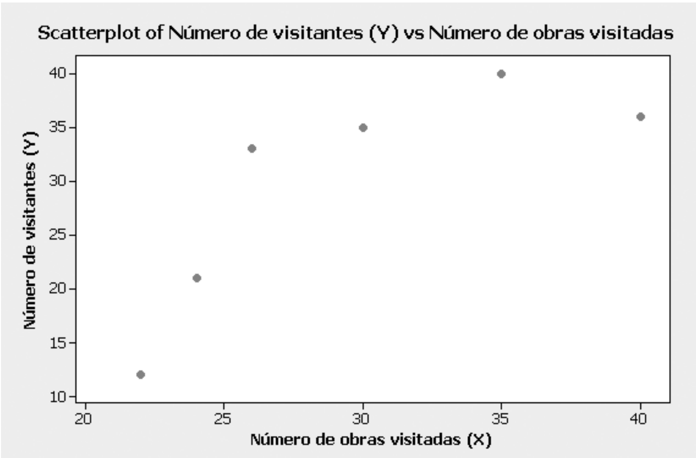
Tabla 5. Número de visitantes a un museo

Número de visitantes (*Y*)	22	24	26	30	35	40
Número de obras visitadas (*X*)	12	21	33	35	40	36

Con estos datos podemos deducir si existe relación entre ambas variables y si las variables están relacionadas establecer el mejor modelo.

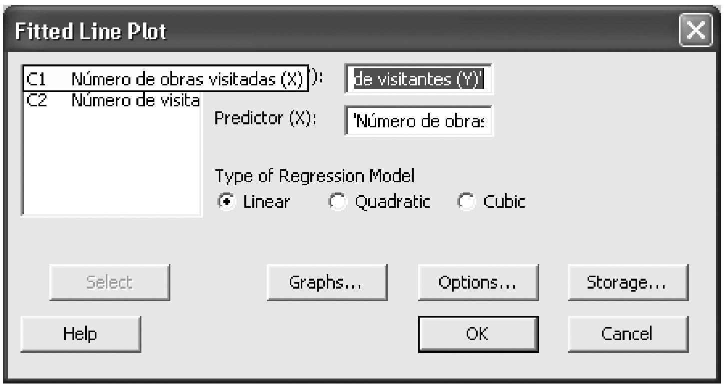
La figura 25 representa el diagrama de dispersión para estos datos. El diagrama de dispersión indica que posiblemente hay una relación curvilínea entre el número de de obras visitadas y el número de visitantes.

Figura 25. Diagrama de dispersión para ejemplo 2. Minitab



Antes de deducir la ecuación curvilínea entre número de obras visitadas y número de visitantes, se realiza el ajuste a un modelo de regresión lineal simple (de primer orden) siguiendo los pasos que muestra la figura 26.

Figura 26. Pasos a seguir para comprobar el modelo lineal



Pasos a seguir

Se sigue la ruta ***Stat*** > ***Regresión*** > ***Fitted Line Plot*** > ***Linear*** y se rellenan los campos en la ventana correspondiente. Seleccionad ***OK*** para obtener el *output* de la figura 27 y 28.

Figura 27. Gráfica de la ecuación de regresión de mínimos cuadrados

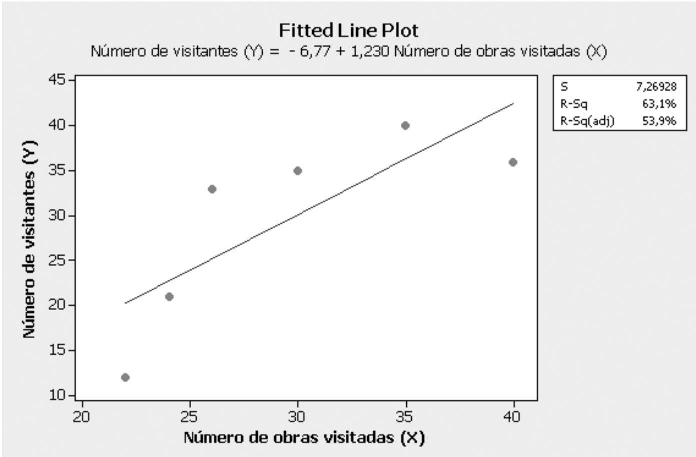


Figura 28. Resultados del análisis de regresión. Modelo lineal simple

Regression Analysis: Número de visitantes (Y) versus Número de obras visitada

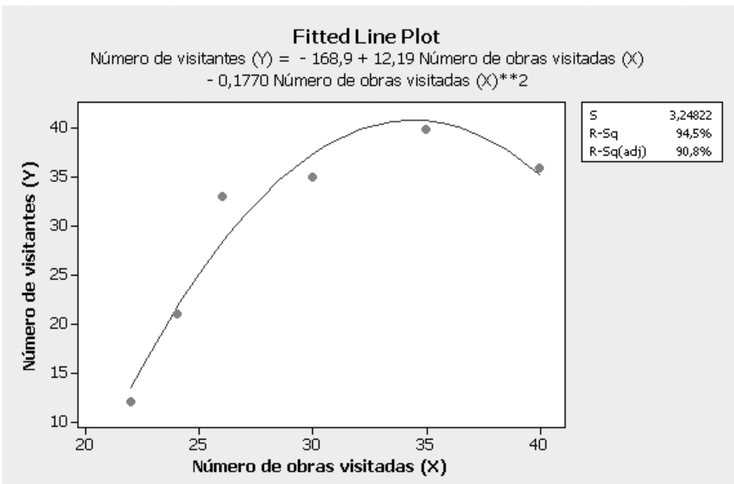
```
The regression equation is
Número de visitantes (Y) = - 6,77 + 1,230 Número de obras
visitadas (X)
S = 7,26928   R-Sq = 63,1%   R-Sq(adj) = 53,9%
Analysis of Variance
Source      DF       SS       MS     F      P
Regression   1  362,130  362,130  6,85  0,059
Error        4  211,370   52,842
Total        5  573,500
```

Observamos que con el modelo lineal se explica un 63,1% de la variabilidad del número de visitantes (R^2 = 63,1%). La ecuación de ajuste es:

Número de visitantes (Y) = –6,77 + 1,230; número de obras visitadas (X)

A continuación se presenta el ajuste del modelo cuadrático y, como se puede ver en la gráfica de la figura 29, los puntos se ajustan mejor a una función no lineal.

Figura 29. Gráfica del ajuste cuadrático



Observamos que el ajuste cuadrático es muy bueno con un valor de R^2= 94,5% que mejora el ajuste lineal. La ecuación de ajuste es:

Número de visitantes (*Y*) = –168,9 + 12,19; número de obras visitadas –0,1770; número de obras visitadas^2

Figura 30. Resultados del análisis de regresión. Modelo cuadrático

Polynomial Regression Analysis: Número de visita versus Número de obras

```
The regression equation is
Número de visitantes (Y) = - 168,9 + 12,19 Número de obras
visitadas (X) - 0,1770 Número de obras visitadas (X)**2



S = 3,24822   R-Sq = 94,5%   R-Sq(adj) = 90,8%

Analysis of Variance

Source      DF       SS       MS      F      P
Regression   2  541,847  270,924  25,68  0,013
Error        3   31,653   10,551
Total        5  573,500

Sequential Analysis of Variance
Source     DF       SS      F      P
Linear      1  362,130   6,85  0,059
Quadratic   1  179,717  17,03  0,026
```

A continuación se presenta el ajuste del modelo cúbico:

Figura 31. Gráfica del ajuste cúbico

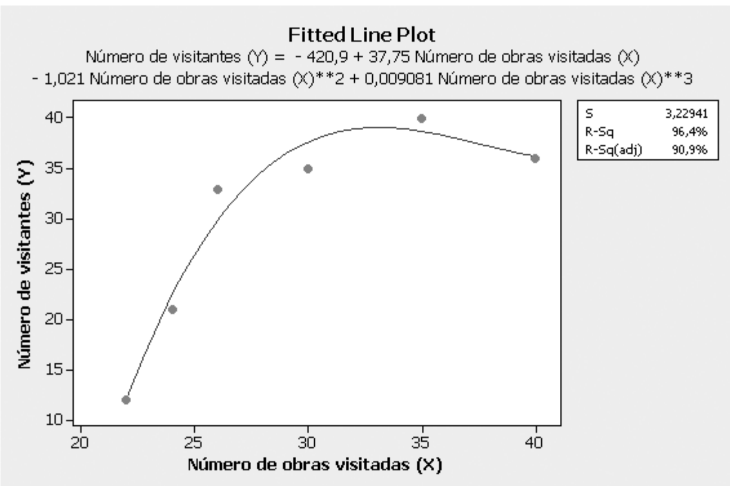


Figura 32. Resultados del análisis de regresión. Modelo cúbico

Polynomial Regression Analysis:
Número de visita versus Número de obras

```
The regression equation is

Número de visitantes (Y) =
- 420,9 + 37,75 Número de obras visitadas (X)
- 1,021 Número de obras visitadas (X)**2
+ 0,009081 Número de obras visitadas (X)**3

S = 3,22941   R-Sq = 96,4%   R-Sq(adj) = 90,9%

Analysis of Variance

Source      DF       SS       MS      F      P
Regression   3  552,642  184,214  17,66  0,054
Error        2   20,858   10,429
Total        5  573,500

Sequential Analysis of Variance

Source     DF       SS      F      P
Linear      1  362,130   6,85  0,059
Quadratic   1  179,717  17,03  0,026
Cubic       1   10,795   1,04  0,416
```

El ajuste al modelo cúbico también es bueno con un valor alto de $R^2 = 96{,}4$ % que mejora el ajuste lineal e iguala al cuadrático.

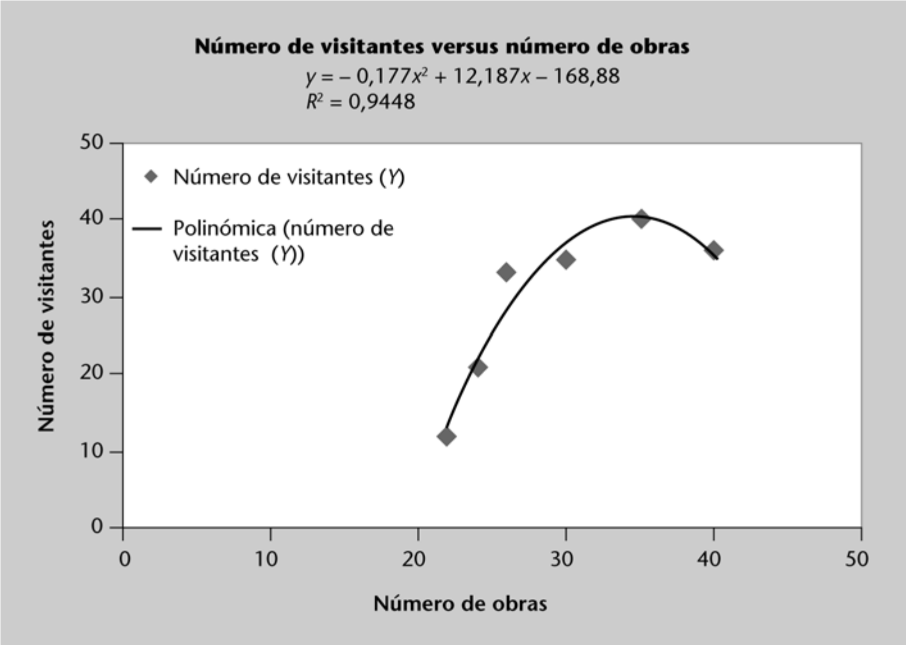
La ecuación de ajuste es:

$$\text{Número de visitantes } (Y) = -420{,}9 + 37{,}75 \text{ Número de obras visitadas } -1{,}021 \text{ Número de obras visitadas}^2 + 0{,}009081 \text{ Número de obras visitadas}^3$$

Analizando la significatividad de los modelos mediante el *p*-valor, el modelo cuadrático por tener el menor *p*-valor (*p*-valor = 0,026) es el más significativo, por lo que se elegiría como mejor ajuste el cuadrático.

La figura 33 muestra el correspondiente *output* que ofrece **Microsoft Excel** del ejemplo 2. "**Número de visitantes a un museo".** Seleccionando la opción Tipo de tendencia poligonal de segundo orden, que coincide con el ajuste cuadrático elegido con Minitab (figuras 29 y 30). La ecuación de ajuste y el valor de R^2 coinciden con las obtenidas con Minitab.

Figura 33. Gráfica del ajuste cuadrático. Excel



3.3. Transformaciones de modelos de regresión no lineales: modelos exponenciales

Algunas relaciones entre variables pueden analizarse mediante modelos exponenciales. Por ejemplo las relaciones entre la variable tiempo (X) y otras variables (Y) como la población, los precios de algunos productos, el número de ordenadores infectados son exponenciales. Los modelos exponenciales de demanda se utilizan mucho en el análisis de conducta del mercado.

Curva en un modelo exponencial

En el modelo lineal se ajusta la nube de puntos a una recta de ecuación:

$$y = a + bx$$

En el modelo exponencial se ajusta a una curva de ecuación:

$$y = ka^x \quad \text{con}\, a > 0,\, k > 0$$

El modelo exponencial es del tipo:

$$y = ka^x \quad \text{con}\, a > 0, k > 0$$

donde k y a son valores constantes.

Para tratar este modelo se realizará una transformación de las variables de manera que el modelo se convierta en lineal.

Si en la ecuación $y = ka^x$ tomamos logaritmos $\ln y = \ln(ka^x)$, obtenemos, por aplicación de las propiedades de los logaritmos:

$$\ln y = \ln k + x \ln a$$

Esta ecuación muestra un modelo lineal entre las variables X y $\ln Y$.

Propiedades de los logaritmos

$$\ln ab = \ln a + \ln b$$
$$\ln a^x = x \ln a$$

Si representamos el diagrama de dispersión de los puntos (x_i, ln y_i) y la nube de puntos presenta una estructura lineal, se puede pensar que entre las variables X e Y hay una relación exponencial.

4. Modelos de regresión múltiple

En el apartado 3.1 hemos presentado el método de regresión simple para obtener una ecuación lineal que predice una variable dependiente o endógena en función de una única variable independiente o exógena: número total de libros vendidos en función del precio. Sin embargo, en muchas situaciones, varias variables independientes influyen conjuntamente en una variable dependiente. La regresión múltiple permite averiguar el efecto simultáneo de varias variables independientes en una variable dependiente utilizando el principio de los mínimos cuadrados.

Existen muchas aplicaciones de la regresión múltiple para dar respuesta a preguntas como las siguientes:

¿En qué medida el precio de un ordenador depende de la velocidad del procesador, de la capacidad del disco duro y de la cantidad de memoria RAM?

¿Cómo relacionar el índice de impacto de una revista científica con el número total de documentos publicados y el número de citas por documento?

¿El sueldo de un titulado depende de la edad, de los años que hace que acabó los estudios, de los años de experiencia en la empresa, etc.?

¿El precio de alquiler de un piso depende de los metros cuadrados de superficie, de la edad de la finca, de la proximidad al centro de la ciudad, etc.?

¿El precio de un coche depende de la potencia del motor, del número de puertas y de multitud de accesorios que puede llevar: airbag, ordenador de viaje, equipo de alta fidelidad volante deportivo, llantas especiales, etc.?

Los métodos para ajustar modelos de regresión múltiple se basan en el mismo principio de mínimos cuadrados explicado en el apartado 3.1.

Nuestro objetivo es aprender a utilizar la regresión múltiple para crear y analizar modelos. Por lo tanto se aprenderá cómo funciona la regresión múltiple y algunas directrices para interpretarla. Comprendiendo perfectamente la regresión múltiple, es posible resolver una amplia variedad de problemas aplicados. Este estudio de los métodos de regresión múltiple es paralelo al de regresión simple. El primer paso para desarrollar un modelo consiste en la selección de las variables y de la forma del modelo. A continuación, estudiamos el método de mínimos cuadrados y analizamos la variabilidad para identificar los efectos de cada una de las variables de predicción.

Después estudiamos la estimación, los intervalos de confianza y el contraste de hipótesis. Utilizamos aplicaciones informáticas para indicar cómo se aplica la teoría a problemas reales.

Desarrollo del modelo

Cuando se aplica la regresión múltiple, se construye un modelo para explicar la variabilidad de la variable dependiente. Para ello hay que incluir las influencias simultáneas e individuales de varias variables independientes. Se supone, por ejemplo, que se quiere desarrollar un modelo que prediga el precio de las impresoras láser que desea liquidar una empresa. Un estudio inicial indicaba que el precio estaba relacionado con el número de páginas por minuto que la impresora es capaz de imprimir y los años de antigüedad de la impresora en cuestión. Eso llevaría a especificar el siguiente modelo de regresión múltiple con dos variables independientes.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

donde:

Y = precio en euros
X_1 = número de páginas impresas por minuto
X_2 = años de antigüedad de la impresora

La tabla 6 contiene 12 observaciones de estas variables. Utilizaremos estos datos para desarrollar el modelo lineal que prediga el precio de las impresoras en función del número de páginas impresas por minuto y de los años de antigüedad de la impresora.

Tabla 6. Datos del ejemplo “Estudio sobre el precio de impresoras láser en función de su velocidad de impresión y la antigüedad del modelo”.

X_1	6	6	6	6	8	8	8	8	12	12	12	12
X_2	6	4	2	0	6	4	2	0	6	4	2	0
Y	466	418	434	487	516	462	475	501	594	553	551	589

Nota

En el caso general emplearemos k para representar el número de variables independientes.

Pero antes de poder estimar el modelo es necesario desarrollar y comprender el método de regresión múltiple.

El modelo de regresión múltiple es

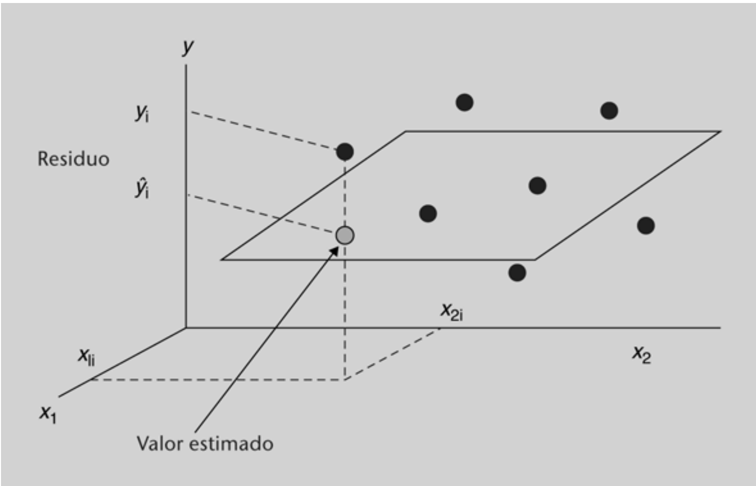
$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \varepsilon_i$$

Donde β_0, β_1, β_2,... , β_k son los coeficientes de las variables independientes o exógenas y ε (letra griega épsilon) es el error o residuo y es una variable alea-

toria. Más adelante describiremos todos los supuestos del modelo para el modelo de regresión múltiple y para ε.

Los coeficientes en general no se conocen y se deben determinar a partir de los datos de una muestra y empleándose el **método de mínimos cuadrados** para llegar a la ecuación estimada de regresión que más se aproxima a la relación lineal entre las variables independientes y dependiente. El procedimiento es similar al utilizado en la regresión simple. En la regresión múltiple el mejor ajuste es un hiperplano en espacio *n*-dimensional (espacio tridimensional en el caso de dos variables independientes, figura 34).

Figura 34. Gráfica de la ecuación de regresión, para el análisis de regresión múltiple con dos variables independientes



Los valores estimados de la variable dependiente se calculan con la ecuación estimada de regresión múltiple:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \ldots + \hat{\beta}_k x_k$$

Donde $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_k$ son los valores de los estimadores de los parámetros o coeficientes de la ecuación de regresión múltiple, la deducción de estos coeficientes requiere el empleo del álgebra de matrices y se sale del propósito de este texto. Así, al describir la regresión múltiple lo enfocaremos hacia cómo se pueden emplear los programas informáticos de cálculo para obtener la ecuación estimada de regresión y otros resultados y su interpretación, y no hacia cómo hacer los cálculos de la regresión múltiple.

Criterio de mínimos cuadrados

$$\min \sum_{i=1}^{n} \left(y_i - \hat{y}_i\right)^2$$

Donde:

y_i = valor observado de la variable dependiente en la *i*-ésima observación.

$\hat{y}_i$ = valor estimado de la variable dependiente en la *i*-ésima observación.

Considerando de nuevo el modelo de regresión con dos variables independientes del ejemplo 3. “Estudio sobre el precio de impresoras láser en función de su velocidad de impresión y la antigüedad del modelo”. Utilizando los datos de la tabla 6 se ha estimado un modelo de regresión múltiple, que se observa en la salida Minitab de la figura 35.

Figura 35. Resultados del ejemplo 3 del análisis de regresión múltiple para dos variables independientes

```
Regression Analysis: Y versus X1; X2

The regression equation is
Y = 330 + 20,2 X1 - 0,50 X2

Predictor    Coef  SE Coef      T      P
Constant   330,38    29,40  11,24  0,000
X1         20,187    3,056   6,61  0,000
X2         -0,500    3,410  -0,15  0,887

S = 26,4100   R-Sq = 82,9%   R-Sq(adj) = 79,1%
```

Pasos a seguir

Para estimar el modelo de regresión múltiple introducimos los datos en Minitab para calcular el modelo.

Se sigue la ruta ***Stat*** > ***Regression*** > ***Regression*** y se rellenan los campos en la ventana correspondiente. Se selecciona ***OK*** para obtener el análisis de regresión.

Los coeficientes estimados se identifican en la salida de los programas informáticos

La ecuación de regresión múltiple es: $Y = 330 + 20{,}2\ X1 - 0{,}50\ X2$

La interpretación de los coeficientes es la siguiente:

- Coeficiente de $X1$ (20,2 euros): sería el aumento del precio de la impresora cuando aumenta en una unidad el número de páginas por minuto que imprime, cuando las demás variables independientes se mantienen constantes (en este caso $X2$, la antigüedad no varía).

- Coeficiente $X2$ (–0,50 euros): sería la disminución del precio por cada año más de antigüedad de la impresora, cuando $X1$ permanece constante (el número de páginas por minuto no varía).

- Término independiente (330): no tiene mucho sentido interpretarlo en este caso ya que representaría el precio de una impresora que no puede imprimir ninguna página.

El coeficiente de determinación múltiple

En la regresión lineal simple vimos que la suma total de cuadrados se puede descomponer en dos componentes: la suma de cuadrados debida a la regresión y la suma de cuadrados debida al error. Este mismo procedimiento se aplica a la suma de cuadrados de la regresión múltiple. El coeficiente de determinación múltiple mide la bondad de ajuste para la ecuación de regresión múltiple. Este coeficiente se calcula como sigue:

Coeficiente de determinación R^2

El coeficiente de determinación R^2 en Minitab se designa como ***R-sq***.

$$R^2 = \frac{SSR}{SST}$$

Se puede interpretar como la proporción de variabilidad de la variable dependiente que se puede explicar con la ecuación de regresión múltiple. Cuando se

multiplica por cien, se interpreta como la variación porcentual de y que se explica con la ecuación de regresión.

En general, R^2 aumenta cuando se añaden variables independientes (variables explicativas o predictoras) al modelo. Si se añade una variable al modelo, R^2 se hace mayor (o permanece igual), aun cuando esa variable no sea estadísticamente significativa. El **coeficiente de determinación corregido** o **adjusted *R-sq*** elimina el efecto que se produce sobre el *R-sq* cuando se aumenta el número de variables independientes.

El **coeficiente de correlación múltiple** se define como la raíz cuadrada positiva del *R-sq*. Este coeficiente nos proporciona la correlación existente entre la variable dependiente (respuesta) y una nueva variable formada por la combinación lineal de los predictores.

Continuando con el **ejemplo 3. "Estudio sobre el precio de impresoras láser en función de su velocidad de impresión y la antigüedad del modelo"**, interpretaremos el resultado del coeficiente de determinación *R-Sq* = 82,9% (figura 35). Significa que el 82,9% de la variabilidad en el precio de impresoras láser se explica con la ecuación de regresión múltiple, con el número de páginas que imprime por minuto y los años de antigüedad. La figura 35 muestra que el valor *R-Sq* (adj) = 79,1%, significa que si se agregase una variable independiente (predictora) el valor de R^2 no aumentaría.

Supuestos del modelo

Los supuestos acerca del término del error ε, en el modelo de regresión múltiple, son similares a los del modelo de regresión lineal simple.

Por simplicidad, consideraremos un modelo de regresión con sólo dos variables explicativas ($X1$ y $X2$). La ecuación de regresión múltiple, con dos variables independientes será:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

donde los β_i representan coeficientes reales y ε representa el error aleatorio.

1) El error es una variable aleatoria cuyo valor medio u esperado es cero; esto es $E(\varepsilon) = 0$.

2) Para todos los valores de $X1$ y $X2$, los valores de Y (o, alternativamente, los valores de (ε) muestran varianza constante σ^2.

3) Para cada valor de $X1$ y $X2$, la distribución de Y (o, alternativamente, la de ε) es aproximadamente normal.

4) Los valores de *Y* obtenidos (o, alternativamente, los de ε) son independientes.

Hay toda una serie de gráficos que nos pueden ayudar a analizar los resultados de una regresión lineal múltiple y a comprobar si se cumplen o no los supuestos anteriores:

1) Un gráfico de la variable dependiente frente a los valores estimados por el modelo nos ayudará a comprobar visualmente la bondad del ajuste.

2) Representando los residuos frente a los valores estimados podremos comprobar la variabilidad vertical en los datos. Ello nos permitirá saber si se cumple el supuesto de varianza constante.

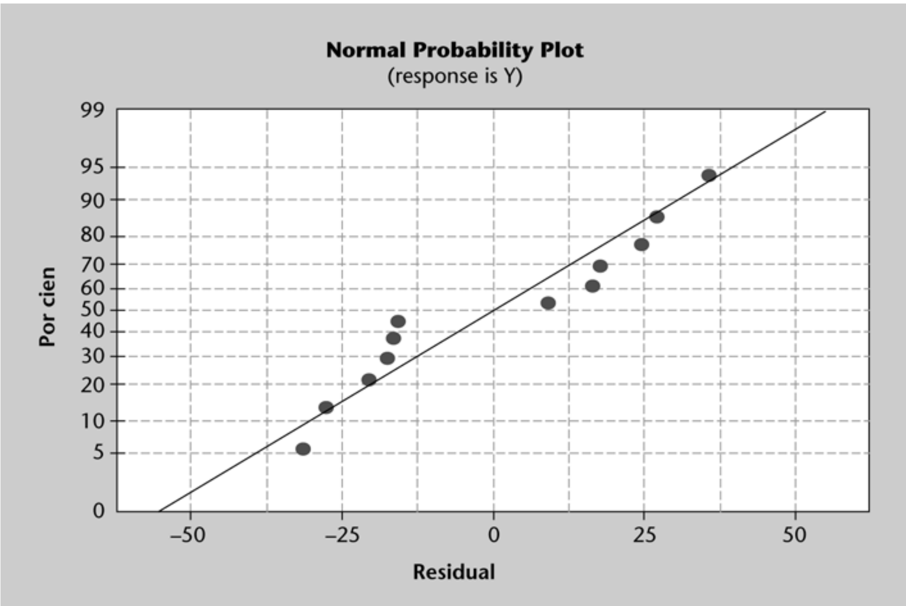
3) Un gráfico de residuos frente a cada una de las variables explicativas puede revelar problemas adicionales que no se hayan detectado en el gráfico anterior.

4) Para comprobar la hipótesis de normalidad suele ser conveniente realizar un test y un gráfico de normalidad para los residuos.

En el ejemplo se comprueba si se cumplen los supuestos del modelo utilizado.

En la gráfica de la figura 36 podemos comprobar que los residuos siguen una distribución aproximadamente normal, ya que los puntos se acercan bastante a una recta.

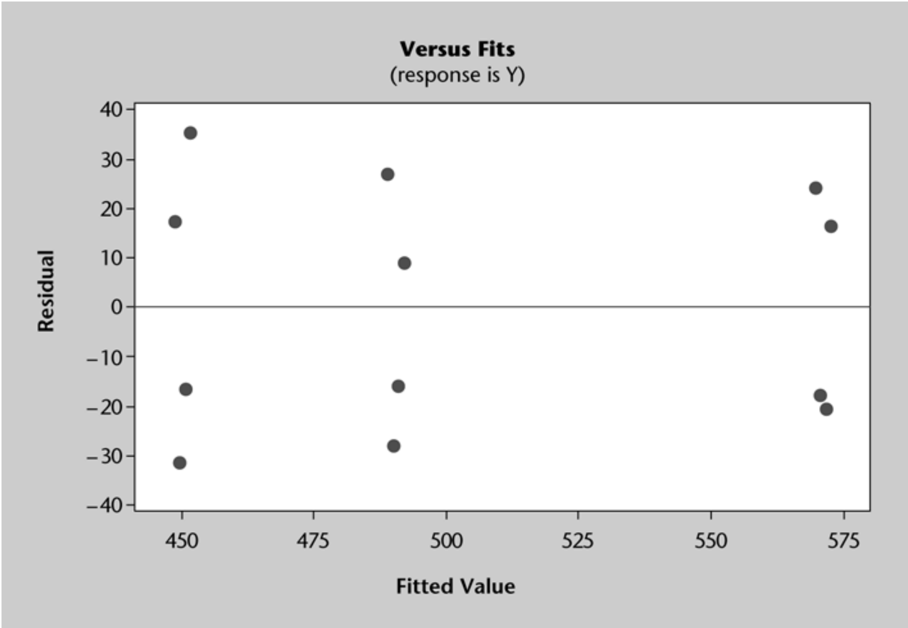
Figura 36. Gráfica de probabilidad normal



La figura 37 presenta el gráfico de los valores residuales frente a los valores estimados. Los residuos se distribuyen aleatoriamente, no presenta ningún tipo de estructura y podemos concluir que es válido el modelo lineal múltiple. También observamos en este gráfico que las varianzas de los residuos son constantes. El procedimiento y la interpretación de los supuestos se explica-

ron en el apartado 3.1. (modelos de regresión lineal simple) y son iguales a los correspondientes de regresión múltiple.

Figura 37. Gráfica de los residuos en función de los valores estimados



Pruebas de significación

Las pruebas de significación que empleamos en la regresión lineal fueron una prueba *t* y una prueba *F*. En ese caso, ambas pruebas dan como resultado la misma conclusión: si se rechaza la hipótesis nula, la conclusión es que $\beta_1 \neq 0$. En la regresión múltiple la prueba *t* y *F* tienen distintas finalidades.

La prueba *F* se usa para determinar si hay una relación significativa entre la variable dependiente y el conjunto de todas las variables independientes. En estas condiciones se le llama **prueba de significación global.**

La prueba *t* se aplica para determinar si cada una de las variables independientes tiene significado. Se hace una prueba *t* por separado para cada variable independiente en el modelo y a cada una de estas pruebas se le llama **prueba de significación individual.**

Prueba *F* o análisis de la varianza en regresión lineal

Las hipótesis para la prueba *F* implican los parámetros del modelo de regresión múltiple:

Hipótesis nula: H_0: $\beta_1 = \beta_2 = \ldots = \beta_k = 0$

Hipótesis alternativa: H_1: uno o más de los parámetros no es igual a cero (al menos un parámetro es $\neq 0$). Debemos fijar el nivel se significación α.

Si se rechaza H_0 tendremos suficiente evidencia estadística para concluir que uno o más de los parámetros no es igual a cero y que la relación general entre y y el conjunto de variables independientes $x_1, x_2,.....x_k$ es significativa. Sin embargo, si no podemos rechazar H_0, no tenemos la evidencia suficiente para llegar a la conclusión de que la relación es significativa.

Para realizar el contraste debemos calcular el estadístico de contraste *F*. **El estadístico *F* es una variable aleatoria que se comporta según una distribución *F*-Snedecor con *k* grados de libertad en el numerador (*DF-Regresión*) y *n*-*k*-1 grados de libertad en el denominador (*DF- Error*).** Donde *k* son los grados de libertad de la regresión iguales a la cantidad de variables independientes y *n* es el número de observaciones. Así pues, el estadístico de contraste es:

$$F^* = \frac{SSR/k}{SSE\,/\,n-k-1}$$

También podemos definir el estadístico de contraste como el cociente de cuadrados medio (*mean squares*).

Cuadrado medio

Es la suma de cuadrados dividida por los grados de libertad (DF) correspondientes. Esta cantidad se usa en la prueba *F* para determinar si hay diferencias significativas entre medias.

El *cuadrado medio debido a la regresión* o simplemente *regresión del cuadrado medio* se representa por **MSR** (*mean square regression*):

$$MSR = \frac{SSR}{\text{grados de libertad de la regresión}} = \frac{SSR}{k}$$

El cuadrado medio debido a los errores o residuos se llama *cuadrado medio residual* o *cuadrado medio del error* se representa por **MSE** (*mean square residual error*):

$$MSE = \frac{SSE}{\text{grados de libertad del error}} = \frac{SSR}{n-k-1}$$

El valor del estadístico de contraste *F* podemos definirlo como: $F^* = \frac{MSR}{MSE}$

Regla de decisión del contraste de hipótesis

Podemos actuar de dos maneras:

a) A partir del *p*-valor. Este valor es: *p*-valor = $P(F_{\alpha;\, k,n-k-1} > F^*)$, donde F_α es un valor de la distribución *F* con *k* grados de libertad en el numerador y $n–k–1$ grados de libertad en el denominador.

- Si *p*-valor < α se rechaza la hipótesis nula H_0; por tanto, el modelo en conjunto explica de forma significativa la variable *Y*. Es decir, el modelo sí contribuye con información a explicar la variable *Y*.

- Si p-valor $\geq \alpha$ no se rechaza la hipótesis nula H_0; por tanto, no hay una relación significativa. El modelo en conjunto no explica de forma significativa la variable Y.

b) A partir de los valores críticos

- Si $F^* > F_{\alpha;\, k,,n-k-1}$, se rechaza la hipótesis nula H_0
- Si $F^* < F_{\alpha;\, k,,n-k-1}$, no se rechaza la hipótesis nula H_0

Tabla de análisis de varianza

En la primera columna se pone la **fuente de variación**, los elementos del modelo responsables de la variación.

En la segunda columna ponemos la **suma de cuadrados** correspondientes.

En la tercera columna ponemos los grados de libertad correspondientes a las **sumas de cuadrados.**

En la cuarta columna con el nombre de **cuadrados medios** se ponen las sumas de cuadrados divididas por los grados de libertad correspondientes. Sólo para *SSR* y *SSE*.

En la quinta columna ponemos el estadístico de contraste ***F***.

Los cálculos necesarios se pueden resumir en la tabla 7, conocida como **tabla de análisis de la varianza:**

Tabla 7. Análisis de varianza para un modelo de regresión múltiple con *k* variables independientes

Fuente de variación	Suma de cuadrados	Grados de libertad	Cuadrados medios	*F*
Regresión	SSR	k	$MSR = SSR/k$	$F = \dfrac{MSR}{MSE}$
Error	SSE	$n-k-1$	$MSE = SSE/n-k-1$	
Total	SST	n-1		

Aplicaremos la prueba F al ejemplo 3. Con dos variables independientes "número de páginas por minuto ($X1$)" y "antigüedad de la impresora ($X2$)".

Las hipótesis se formulan como sigue:

$$H_0\text{: } \beta_1 = \beta_2 = 0$$

$$H_1\text{: } \beta_1 \text{ y/o } \beta_2 \text{ no es igual a cero}$$

Fijamos un nivel de significación del 5% (α = 0,05).

La figura 38 muestra los resultados del modelo de regresión múltiple, en la parte de resultados correspondiente al análisis de varianza.

Figura 38. Resultados obtenidos con Minitab.Tabla de análisis de varianza

```
Analysis of Variance

Source          DF      SS      MS      F       P
Regression       2   30444   15222  21.82   0.000
Residual Error   9    6277     697
Total           11   36722
```

El valor del estadístico de contraste es F^* = 21,82, el p-valor = 0,000

Como p-valor $< \alpha$, rechazamos la hipótesis nula, por tanto, el modelo **en conjunto** explica de forma significativa la variable Y. Es decir, llegamos a la con-

clusión de que hay una relación significativa entre el precio de la impresora y las dos variables independientes que son número de páginas impresas por minuto ($X1$) y la antigüedad de la impresora ($X2$).

Prueba *t*

Se utiliza para determinar el significado de cada uno de los parámetros individuales. Las hipótesis para la prueba *t* implican los parámetros del modelo de regresión múltiple, se hace un contraste para cada parámetro β:

Hipótesis nula: H_0: $\beta_1 = \beta_2 = \ldots = \beta_k = 0$

Hipótesis alternativa: H_1: uno o más de los parámetros no es igual a cero (al menos un parámetro es ≠ 0). Debemos fijar el nivel se significación α.

El estadístico de contraste es:

$$t^* = \frac{\hat{\beta}_i}{S_{\hat{\beta}_i}}$$

Sigue una distribución *t* de Student con n-k-1 grados de libertad

Regla de decisión del contraste de hipótesis

Podemos actuar de dos maneras:

a) A partir del p-valor. Este valor es: $p = 2P(t_{n-k-1} > | t^*|)$.

- Si $p < \alpha$ se rechaza la hipótesis nula H_0; se rechaza la hipótesis nula H_0; por tanto, hay una relación lineal entre la variable X_i e Y. Por consiguiente, dicha variable debe permanecer en el modelo.

- Si $p \geq \alpha$ no se rechaza la hipótesis nula H_0; por tanto, no hay una relación lineal entre la correspondiente variable X_i e Y. Decimos que la variable implicada X_i es no explicativa y podemos eliminarla del modelo.

b) A partir de los valores críticos $\pm t\alpha/2, n-k-1$, de manera que:

- Si $|t^*| > t\alpha/2, n-k-1$, se rechaza la hipótesis nula H_0; por tanto, la variable es significativa.

- Si $|t^*| > t\alpha/2, n-k-1$, no se rechaza la hipótesis nula H_0; por tanto, la variable no es significativa. Decimos que la variable implicada X_i no es explicativa.

Si la prueba *F* del ejemplo (figura 38) ha mostrado que la relación múltiple tiene significado, se puede hacer una prueba *t* para determinar el significado de cada uno de los parámetros individuales. El nivel de significación es $\alpha = 0,05$. Obsérvese que los valores de los estadísticos *t* aparecen en la figura 39. Los *p*-valores de los contrastes individuales son para el contraste de β_1 el *p*-valor = 0,000 y para β_2, *p*-valor = 0,887.

Figura 39. Resultados obtenidos con Minitab

Predictor	Coef	SE Coef	T	P	VIF
Constant	330.38	29.40	11.24	0.000	
X1	20.187	3.056	6.61	0.000	1.000
X2	-0.500	3.410	-0.15	0.887	1.000

Interpretamos el contraste para el parámetro β_1, la H_0: $\beta_1 = 0$, H_1: $\beta_1 \neq 0$. Como $0,000 < 0,05$ se rechaza H_0, y, por tanto, la variable *X*1 (número de páginas impresas por minuto) es significativa.

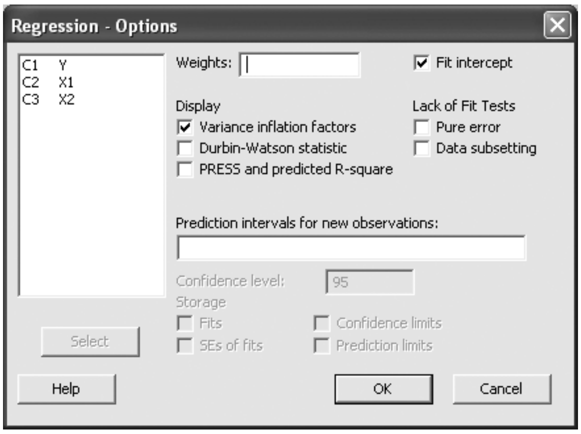
El contraste para el parámetro β_2, la H_0: $\beta_2 = 0$, H_1: $\beta_2 \neq 0$. Como $0,887 > 0,05$ no podemos rechazar H_0, por lo que la variable *X*2 (antigüedad) no es significativa y podríamos eliminarla del modelo porque no influye significativamente en el precio.

El problema de la multicolinealidad

En los problemas de regresión lineal múltiple esperamos encontrar dependencia entre la variable *Y* y las variables explicativas *X*1, *X*2, ..., *Xk*, pero en algunos problemas de regresión podemos tener también algún tipo de dependencia entre algunas de las variables *Xj*. En este caso tenemos información redundante en el modelo. Este fenómeno se llama **multicolinealidad** y suele ser bastante frecuente en los modelos de regresión lineal múltiple.

El término **multicolinealidad** en análisis de regresión múltiple indica la correlación entre variables independientes. La multicolinealidad puede tener efectos muy importantes en las estimaciones de los coeficientes de la regresión y, por tanto, sobre las posteriores aplicaciones del modelo estimado. Cuando las variables independientes están muy correlacionadas no es posible determinar el efecto por separado de una de ellas sobre la variable dependiente. Cuando existe multicolinealidad, los resultados de los contrastes de hipótesis sobre el modelo conjunto y los resultados de los contrastes individuales son aparentemente contradictorios, pero realmente no lo son. Este efecto lo veremos en el ejemplo propuesto (figura 40). Minitab dispone de una opción, llamada **Variance Inflation Factors** (VIF), que nos permite identificar la multicolinealidad entre los predictores del modelo. La figura 40 indica los pasos a seguir.

Figura 40. Pasos a seguir para identificar la multicolinealidad



Pasos a seguir

Se sigue la ruta ***Stat*** > ***Regression*** > ***Regression*** > ***Options*** y se rellenan los campos en la ventana correspondiente. Seleccionad ***OK***.

Ahora la figura 41 de los resultados del análisis de regresión múltiple contiene los valores VIF. Cada coeficiente VIF es de 1,000. Estos valores son bajos, lo que indica que las variables independientes no están correlacionadas. Dado que estos valores indican que el grado de colinearidad es bajo. No existe multicolienalidad en el modelo propuesto.

Figura 41. Resultados del ejemplo 3 del análisis de regresión múltiple, que incluye los *Variance Inflation Factors* (VIF) o factores de inflación de la varianza

Regression Analysis: Y versus X1; X2

```
The regression equation is
Y = 330 + 20.2 X1 - 0.50 X2


Predictor     Coef  SE Coef      T      P    VIF
Constant    330.38    29.40  11.24  0.000
X1          20.187    3.056   6.61  0.000  1.000
X2          -0.500    3.410  -0.15  0.887  1.000


S = 26.4100   R-Sq = 82.9%   R-Sq(adj) = 79.1%
```

Usando **Microsoft Excel para obtener el análisis de regresión del ejemplo 3. "Estudio sobre el precio de impresoras láser en función de su velocidad de impresión y la antigüedad del modelo".**

Pasos a seguir

Para efectuar la regresión múltiple con **MS Excel**, una vez introducidos los datos en la hoja de cálculo se sigue la siguiente ruta: clic en ***Herramientas*** > ***Análisis de datos*** > ***Regresión*** > ***OK***.

A continuación se seleccionan los rangos de datos de las variables.

La tabla 8 muestra el correspondiente *output* que ofrece **Microsoft Excel**.

Tabla 8. Resultados del análisis de regresión del ejemplo 3. Estudio sobre el precio de impresoras láser en función de su velocidad de impresión y la antigüedad del modelo. Excel

	B	C	D	E	F	G	H
1	Resumen						
2							
3	*Estadísticas de la regresión*						
4	Coeficiente de correlación múltiple	0,910524728228339					
5	Coeficiente de determinación R^2	0,829055280715291					
6	R^2 ajustado	0,791067565318689					
7	Error típico	26.40996235					
8	Observaciones	12					
9							
10	ANÁLISIS DE VARIANZA						
11		*Grados de libertad*	*Suma de cuadrados*	*Promedio de los cuadrados*	*F*	*Valor crítico de F*	
12	Regresión	2	30444.29167	15222.14583	21.82429957	0.000353062	
13	Residuos	9	6277.375	697.4861111			
14	Total	11	36721.66667				
15							
16		*Coeficientes*	*Error típico*	*Estadístico t*	*Probabilidad*	*Inferior 95%*	*Superior 95%*
17	Intercepción	330.375	29.40041791	11.23708517	1.34464E-06	263.8666342	396.8833658
18	X1	20.1875	3.056359247	6.605080872	9.86968E-05	13.27353505	27.10146495
19	X2	-0.5	3.409511478	-0.146648575	0.886641778	-8.212850796	7.212850796

Resumen

En este módulo hemos introducido conceptos de relaciones funcionales y estadísticas, así como el de variables dependientes y el de variables independientes. Hemos comentado la construcción de un diagrama de dispersión como paso inicial a la hora de buscar algún tipo de relación entre dos variables. Si el diagrama muestra una estructura lineal, entonces se buscará la recta que mejor se ajusta a las observaciones. Hemos puesto de manifiesto la importancia de interpretar correctamente los coeficientes de la recta. También hemos visto cómo se debe utilizar la recta de regresión para realizar predicciones. Hemos introducido una medida numérica de la bondad de ajuste. Esta medida se obtiene con el coeficiente de determinación, discutiendo los valores que puede tomar. Finalmente, hemos comentado la importancia de analizar los residuos para hacer un diagnóstico del modelo lineal obtenido.

En este módulo de regresión lineal simple hemos considerado que las observaciones sobre dos variables X e Y son una muestra aleatoria de una población y que se utilizan para extraer algunas conclusiones del comportamiento de las variables sobre la población, y para ello hemos visto cómo hacer inferencia sobre la pendiente de la recta obtenida a partir de la muestra y cómo hacer un contraste de hipótesis para decidir si la variable X explica realmente el comportamiento de la variable Y. También hemos comentado algunas las relaciones no lineales y la manera en que se puede transformar en una lineal.

Hemos tratado la regresión lineal múltiple como una generalización del modelo de regresión lineal simple en aquellos casos en los que se tiene más de una variable explicativa. Finalmente, hemos visto cómo hacer inferencia sobre los coeficientes de regresión obtenidos a partir de la muestra, cómo hacer un contraste de hipótesis para cada uno de los coeficientes obtenidos para decidir si las variables independientes explican realmente el comportamiento de la variable dependiente o se puede prescindir de alguna de ellas. También hemos realizado un contraste conjunto del modelo. Finalmente, hemos presentado el posible problema de multicolinealidad que puede aparecer y que es debido a la relación entre algunas de las variables explicativas que supuestamente son independientes.

Ejercicios de autoevaluación

1) Los precios de una pantalla TFT de una conocida marca son los siguientes:

Tamaño (pulgadas)	15	17	19	24
Precio (euros)	251	301	357	556

Calculad la recta de regresión para explicar el precio a partir del tamaño.

2) Con los datos de la cuestión anterior queremos decidir si se trata de un buen modelo. ¿Qué método proponéis para determinar si se ajusta bien? ¿Qué podemos decir del caso concreto del ejemplo anterior?

3) Consideramos un modelo lineal para explicar el rendimiento de un sistema informático (variable Y) en relación con el número de *buffers* y el número de procesadores (variables $X1$ y $X2$ respectivamente). Se obtiene el modelo $Y = -3{,}20 + 2X1 + 0{,}0845X2$ con un coeficiente de determinación de 0,99. ¿Se trata de un buen modelo? ¿Cuál será el rendimiento esperado si tenemos 1 *buffer* y 1 procesador? Comentad si este valor os parece lógico y si puede relacionarse con la bondad del modelo.

4) La empresa Ibérica editores tiene que decidir si firma o no un contrato de mantenimiento para su nuevo sistema de procesamiento de palabras. Los directivos creen que el gasto de mantenimiento debe estar relacionado con el uso y han reunido la información que vemos en la tabla siguiente sobre el uso semanal, en horas, y el gasto anual de mantenimiento (cientos de euros).

Uso semanal (horas)	**Gastos anuales de mantenimiento**
13	17,0
10	22,0
20	30,0
28	37,0
32	47,0
17	30,5
24	32,5
31	39,0
40	51,5
38	40,0

a) Determinad la ecuación de regresión que relaciona el costo anual de mantenimiento con el uso semanal.

b) Probad el significado de la relación obtenida en el apartado *a* al nivel de significación 0,05.

c) Ibérica editores espera usar el procesador de palabras 30 horas semanales. Determinad un intervalo de predicción del 95% para el gasto de la empresa en mantenimiento anual.

d) Si el contrato de mantenimiento cuesta 3.000 euros anuales, ¿recomendaríais firmarlo? ¿Por qué?

5) Una biblioteca pública de una ciudad española ofrece un servicio vía Internet de préstamo de libros a los usuarios. Se quiere estudiar la correlación entre el número de usuarios de esta biblioteca virtual y cuántos de ellos acaban realizando los préstamos.

Los datos de los últimos doce meses son:

Usuarios	296	459	602	798	915	521	362	658	741	892	936	747
Préstamos	155	275	322	582	761	324	221	415	562	628	753	569

a) Determina el coeficiente de correlación entre las dos variables. Calcula y representa la recta de regresión.

b) ¿Qué número de préstamos se esperaría si el número de usuarios aumentase a 1.000?

6) Un experto documentalista necesita saber si la eficiencia de un nuevo programa de búsqueda bibliográfica depende del volumen de los datos entrantes. La eficiencia se mide con el número de peticiones por hora procesadas. Aplicando el programa a distintos volúmenes de datos, obtenemos los resultados siguientes:

Volumen (gigabytes), *X*	6	7	7	8	10	10	15
Peticiones procesadas, *Y*	40	55	50	41	17	26	16

a) Calculad la recta de regresión para explicar las peticiones procesadas por hora a partir del volumen de datos e interpretad los parámetros obtenidos.

b) Cread el gráfico de ajuste a la recta de mínimos cuadrados.

c) Determinad el coeficiente de correlación lineal entre las dos variables e interpretad su significado.

d) Determinad el coeficiente de determinación R^2 e interpretad su significado.

e) Calculad, a partir de la recta anterior, cuántas peticiones podemos esperar para un volumen de datos de 12 gigabytes.

f) Realizad el contraste de hipótesis sobre la pendiente. ¿Podemos afirmar a un nivel de significación de 0,05 que la pendiente de la recta es cero?

Solucionario

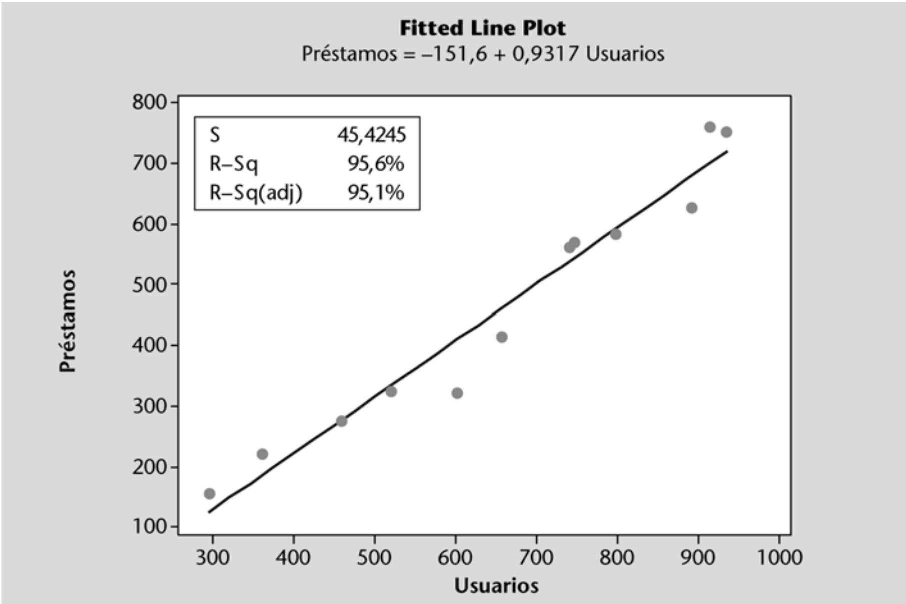
1) Precio = –279,11 + 34,42 · tamaño.

2) Para estudiar la calidad del ajuste, se calcula el coeficiente de correlación muestral $r = 0{,}994$

3) Es un buen modelo ya que el coeficiente de determinación es muy cercano a 1. El rendimiento, si tenemos un *buffer* y un procesador sería: $Y = –3{,}20 + 2 \cdot 1 + 0{,}0845 \cdot 1 = –1{,}1155$. Este valor no tiene sentido, ya que el rendimiento no puede ser negativo. De todas las maneras, este hecho no es contradictorio con tener un buen modelo ya que estamos fuera del intervalo donde la regresión funciona.

4)
a) $\hat{y} = 10{,}5 + 0{,}953x$.
b) Relación significativa; *p*-valor = 0,000.
c) $[2.874; 54.952]$ euros.
d) Sí, la probabilidad de encontrar el gasto de mantenimiento dentro del intervalo de confianza es del 95%.

5)
a) $r = 0{,}978$.



b) $–151{,}6 + 0{,}9317 \times 1.000 \simeq 780$ préstamos

6)
a)

Regression Analysis: Peticiones procesadas versus Volumen (gigabytes)

```
The regression equation is
Peticiones procesadas = 72,29 - 4,143 Volumen (gigabytes)
S = 9,90815   R-Sq = 66,2%   R-Sq(adj) = 59,4%

Analysis of Variance
Source      DF       SS       MS     F      P
Regression   1   961,14  961,143  9,79  0,026
Error        5   490,86   98,171
Total        6  1452,00
```

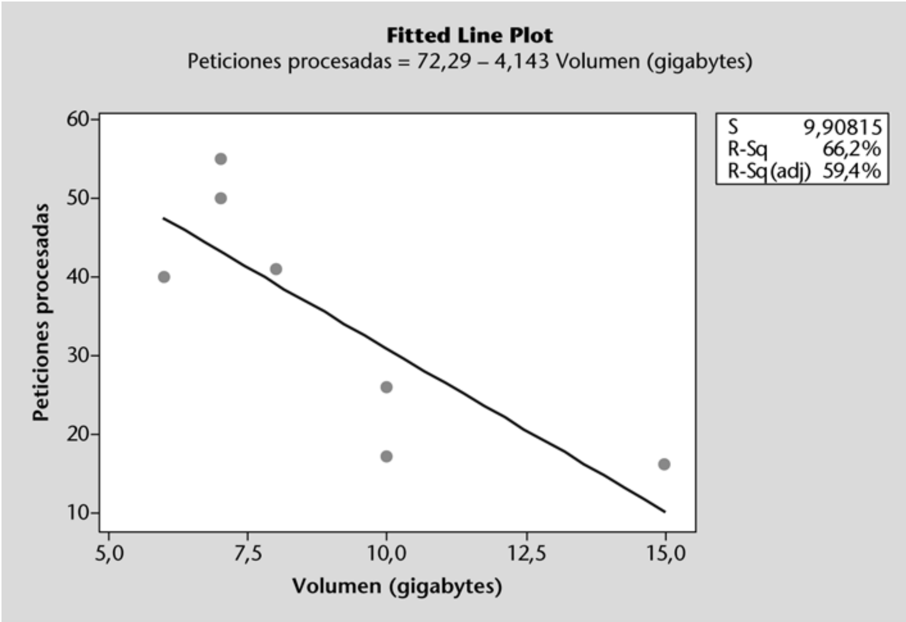
La recta de regresión será:

Peticiones procesadas = 72,29 – 4,143 volumen (gigabytes).

La ordenada en el origen: 72,29; en este caso su significado no tiene ningún sentido.

La pendiente de la recta: –4,143; es negativa: indica que, por cada unidad de volumen de datos (gigabytes) que aumenten los datos entrantes, el número de peticiones procesadas disminuye en 4,143 unidades.

b) El gráfico de ajuste a la recta de mínimos cuadrados es:



c)

Correlations: Volumen (gigabytes); Peticiones procesadas

```
Pearson correlation of Volumen (gigabytes) and Peticiones
procesadas = -0,814
 P-Value = 0,026
```

El coeficiente de correlación $r = -0{,}814$ nos indica que hay una correlación alta negativa entre volumen de datos entrantes y el número de peticiones procesadas.

d) El coeficiente de determinación *R-Sq* es el 66,2%. Esto quiere decir que nuestro modelo lineal explica el 66,2% del comportamiento de la variable Y (en este caso, número de peticiones procesadas).

e) Con 12 gigabytes, habrá $72{,}3 - 4{,}14 \cdot 12 = 22{,}57$ peticiones.

f) En el *output* anterior podemos ver que el p-valor asociado al contraste de hipótesis anterior es 0,026. Como este valor es menor que $\alpha = 0{,}05$, debemos rechazar la hipótesis nula; es decir, podemos concluir que la pendiente de la recta es distinta de cero o, lo que es lo mismo, que el coeficiente de correlación poblacional es no nulo (es decir, que ambas variables están correlacionadas y que, por tanto, el modelo estudiado tiene sentido).

Introducción al diseño y análisis de encuestas

Aplicaciones estadísticas a la selección de muestras y al análisis de cuestionarios

Ángel A. Juan y Alicia Vila

PID_00161062

Universitat Oberta de Catalunya
www.uoc.edu

Índice

Introducción

Las encuestas y cuestionarios se han convertido en una herramienta de investigación de uso cotidiano en la llamada “sociedad de la información”. La idea de usar datos provenientes de una muestra –compuesta por un número relativamente pequeño de elementos– para obtener información sobre toda una población es utilizada a diario por los medios de comunicación, ya sea prensa escrita, televisión, radio o incluso Internet.

En efecto, las encuestas y los cuestionarios se usan para sondear el estado de opinión de los potenciales votantes de unas elecciones, para conocer el potencial interés de nuevos bienes o servicios en el mercado, para predecir la aceptación que tendrán determinadas decisiones gubernamentales o estratégicas, para conocer mejor a los miembros de una comunidad, para detectar demandas potenciales de los consumidores que no están siendo satisfechas, etc. En investigación, además, las técnicas basadas en el uso de encuestas y cuestionarios representan probablemente la herramienta de investigación social más común en artículos y publicaciones científicas.

Sin embargo, el paso de datos muestrales a información sobre la población no es trivial, ya que requiere de todo un proceso metódico que incluye el diseño de las preguntas (para evitar introducir sesgos innecesarios en las mismas), el diseño de la muestra (para minimizar en lo posible el error muestral), la realización de la encuesta y el análisis de los resultados. En muchas ocasiones este proceso se hace demasiado a la ligera y de forma poco rigurosa, con lo que los resultados que se obtienen son poco fiables y nada creíbles desde un punto de vista científico. En este módulo se presentan y discuten los conceptos básicos de estas técnicas, desde las claves de un buen cuestionario y de un buen diseño muestral hasta ejemplos de cómo pueden aplicarse las técnicas estadísticas trabajadas durante el curso para representar numérica y gráficamente la información obtenida sobre la población.

Objetivos

Los objetivos docentes que se pretenden alcanzar con este módulo son los siguientes:

1. Entender la importancia de las encuestas y los cuestionarios en la sociedad de la información.

2. Conocer los aspectos clave a considerar cuando se elaboran las preguntas de un cuestionario.

3. Conocer los tipos de escalas más habituales en los cuestionarios, así como el tipo de datos que produce cada una de ellas.

4. Introducirse en los tipos de muestreo más habituales en los estudios de encuestas, en particular: el muestreo aleatorio simple, el muestreo sistemático, el muestreo por estratos y el muestreo por conglomerados.

5. Saber calcular estimaciones puntuales y por intervalos para diversos parámetros poblacionales según el tipo de muestreo usado.

6. Aprender a usar las técnicas estadísticas trabajadas durante el curso para analizar cuestionarios.

7. Aprender a usar programas estadísticos o de análisis de datos como instrumento básico en la aplicación práctica de los conceptos y técnicas estadísticas.

1. Diseño de cuestionarios

Las técnicas de investigación basadas en el uso de encuestas se aplican a multitud de ámbitos diferentes: en los negocios, en la administración pública, en las ciencias sociales y del comportamiento, en las ciencias de la información y la comunicación, en las ciencias de la salud, en las ciencias políticas, y en cualquier otro ámbito en el que los datos que puedan aportar los usuarios de un servicio o los consumidores de un producto jueguen un papel fundamental. En la Sociedad de la Información, las organizaciones e instituciones hacen un uso intensivo de los datos que explican cómo se comportan los individuos, cuáles son sus gustos y sus necesidades, qué opinión tienen sobre determinados temas, etc. En este contexto, las técnicas de investigación basadas en el uso de encuestas permiten obtener unos datos que, tras su posterior análisis estadístico, proporcionan una valiosa información tanto a los investigadores teóricos de una determinada disciplina como a los responsables de tomar decisiones sobre el funcionamiento de las organizaciones.

En general, se pueden distinguir seis fases secuenciales en el desarrollo de cualquier estudio basado en el uso de encuestas (figura 1): (a) identificación de los temas concretos sobre los que se desea obtener información así como de la población a encuestar, (b) diseño del cuestionario como instrumento para obtener los datos que se necesitan, (c) diseño y selección de una muestra representativa de la población, (d) obtención de los datos mediante el envío del cuestionario a los individuos que componen la muestra, (e) análisis estadístico de las observaciones muestrales a fin de inferir información sobre la población, y (f) elaboración de informes y conclusiones.

Figura 1. Fases en el desarrollo de una encuesta



En este apartado se hará especial énfasis en la fase de diseño del cuestionario, dejando para apartados posteriores otras fases clave en las que las técnicas estadísticas tienen una aportación decisiva, es decir, la fase de diseño y selección de la muestra y la fase de análisis de las observaciones muestrales.

1.1. Elaboración de las preguntas de un cuestionario

Las preguntas que se formulan en un cuestionario constituyen el aspecto más relevante de cualquier encuesta. Para que éstas cumplan su papel de forma efi-

ciente, las preguntas de un cuestionario deben centrarse en aquellos aspectos esenciales sobre los que se desea obtener información. Asimismo, dichas preguntas deben ser lo más breves y claras posibles a fin de facilitar la tarea de las personas encuestadas y maximizar la fiabilidad y validez del cuestionario. Se trata de evitar posibles problemas tales como: interpretaciones erróneas de las preguntas, agotamiento del encuestado o, incluso, rechazo a contestar una parte o la totalidad del cuestionario por la longitud del mismo o el esfuerzo necesario para entender las preguntas y contestarlas. Estas problemáticas podrían introducir sesgos y errores muestrales en los datos, lo que mermaría la fiabilidad y validez de la encuesta y de sus resultados.

Es importante ser cuidadoso en la elaboración de las preguntas a fin de evitar introducir en el cuestionario problemas de **error muestral** –debido al uso de una muestra para estimar parámetros poblacionales– o de **sesgo** (cualquier otro tipo de error en el cuestionario diferente del error muestral): si en la propia formulación de las pregunta se está induciendo al encuestado a responder en un sentido concreto, entonces se está introduciendo un sesgo en el cuestionario; si la formulación de las preguntas es ambigua y da pie a diferentes interpretaciones, entonces se está favoreciendo una excesiva dispersión de las respuestas, lo que incrementa el error muestral. Por tanto, la manera en cómo las preguntas se formulan en un cuestionario es determinante a la hora de evitar introducir patrones de sesgo y error muestral en el mismo. Así, se pueden establecer las siguientes recomendaciones generales a tener presentes cuando se elaboran las preguntas de un cuestionario:

- Criterios de interpretación y respuesta claros: los criterios en los que el encuestado debe basarse para interpretar y contestar a una pregunta deben estar claramente especificados en el cuestionario.

- Preguntas apropiadas al conjunto de individuos que configuran la muestra: todos los encuestados deben poder responder a las preguntas sobre la base de su experiencia o condición personal.

- Uso adecuado de expresiones, ejemplos o alternativas de respuesta: debe evitarse incluir en la pregunta expresiones que inciten a una determinada respuesta, así como ejemplos de posibles respuestas, ya que ello podría inducir a los encuestados a responder de una determinada manera y de este modo introducir un factor de sesgo en las respuestas.

- Nivel de actualidad de las preguntas: no se debería presuponer que el encuestado será capaz de recordar con precisión cuál fue su comportamiento en el pasado o su opinión sobre un tema acontecido hace ya bastante tiempo.

- Preguntas con un nivel de generalización o concreción adecuado: se debería evitar formular preguntas demasiado genéricas o ambiguas que se pue-

dan interpretar de formas muy distintas y cuya respuesta no aporte demasiada información, así como preguntas demasiado específicas que el encuestado no sea capaz de contestar con el nivel de detalle requerido.

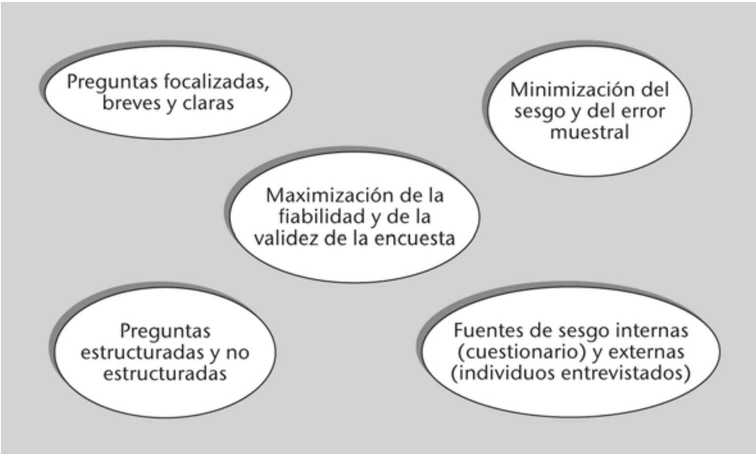
Además de estas fuentes internas de sesgo causadas por el propio instrumento de la encuesta, existen también otras potenciales fuentes de sesgo que no se originan por cómo se han elaborado las preguntas, sino por las condiciones en las que se ha respondido al cuestionario. Conviene conocer y tener presentes estas otras fuentes potenciales de sesgo para evitarlas en lo posible con una correcta elección de las condiciones de la encuesta y, en particular, de la muestra. Así, algunas de estas fuentes externas de sesgo son las siguientes: respuestas que buscan estar en coherencia con lo que es "socialmente deseable" o con lo que el entrevistador espera obtener, respuestas orientadas a dar una buena imagen del encuestado, respuestas con excesiva tendencia a la dicotomía (sí o no, positivo o negativo, etc.) o hacia las opciones extremas, respuestas hostiles excesivamente condicionadas por experiencias negativas recientes, etc.

Existen dos formatos básicos para elaborar preguntas de un cuestionario: las **preguntas abiertas** o no estructuradas son aquellas que permiten al encuestado responder libremente sin estar condicionado por un conjunto de posibles alternativas de respuesta. Por el contrario, las **preguntas estructuradas** o cerradas son aquellas que contienen en la propia pregunta un conjunto de posibles respuestas o categorías a elegir por el encuestado. La preguntas estructuradas son las que habitualmente más se usan en los cuestionarios, ya que además de acotar más claramente el contexto de la información que se espera obtener, suelen ser más fáciles y rápidas de contestar, permiten comparar mejor diferentes grupos de encuestados y, sobre todo, facilitan enormemente el procesado y análisis posterior de los datos.

Cuando se usan preguntas estructuradas es importante elegir bien las categorías o posibles respuestas alternativas de manera que éstas constituyan una lista completa de opciones (incluyendo opciones como "otros" o "no sabe o no contesta" cuando sea necesario) y sean mutuamente excluyentes (a menos que sean de opción múltiple). Por lo que respecta al número de categorías o respuestas alternativas, lo recomendable es que se sitúe entre un mínimo de dos para preguntas dicotómicas y un máximo de seis. Añadir más categorías suele dificultar en exceso la tarea del encuestado. Hay que tener presente, sin embargo, que en caso de duda sobre el nivel de detalle que se quiera ofrecer en las categorías, suele ser preferible optar por la opción con más categorías, puesto que siempre es posible combinar o agregar categorías a posteriori –durante la fase de análisis–, mientras que la operación de desagregar respuestas ya obtenidas en nuevas categorías no suele ser posible sin la consiguiente pérdida de precisión e información.

La figura 2 sintetiza los conceptos clave que se deben tener en cuenta en la elaboración de las preguntas de cualquier cuestionario.

Figura 2. Conceptos clave en la elaboración de las preguntas de un cuestionario



1.2. Uso de escalas en preguntas estructuradas

Las respuestas a preguntas estructuradas consisten, por lo general, en elegir una opción concreta en una lista de categorías posibles. Estas categorías siguen una escala o graduación que puede ser simplemente nominal o bien puede implicar algún tipo de relación ordinal o numérica entre las distintas categorías implicadas:

- **Escalas nominales:** son aquellas en las que las categorías no están asociadas a una relación de orden o de magnitud. Un ejemplo sería una escala en la que las categorías fuesen distintos códigos postales, prefijos telefónicos o identificadores del sexo (“hombre”, “mujer”). Este tipo de escala proporciona datos de tipo nominal que simplemente identifican categorías, por lo que es el más limitado desde el punto de vista de las técnicas estadísticas que se pueden aplicar a las observaciones obtenidas.

- **Escalas ordinales:** son aquellas cuyas categorías siguen una relación de orden o preferencia, aunque no de magnitud, que permite clasificarlas. Un ejemplo sería una escala de tareas secuenciales a realizar en un proceso, en el que la pregunta podría ser elegir aquella tarea que se considere más crítica. Este tipo de escalas posibilita el uso de las llamadas técnicas estadísticas no paramétricas para analizar los datos obtenidos.

- **Escalas de intervalos equidistantes:** son las que asocian una magnitud a cada categoría y en las que el cero no significa ausencia de magnitud. Un ejemplo sería una escala graduada del 1 al 7 para representar niveles de importancia. Esta escala permite el uso de técnicas de inferencia estadística, por lo que resulta altamente recomendable.

- **Escalas de ratio:** son las que asocian una magnitud a cada categoría y en las que el cero representa ausencia de magnitud. Un ejemplo sería una escala graduada del 0 al 50 para indicar la distancia en kilómetros recorrida por el encuestado para acudir a su lugar de trabajo. Al igual que ocurría con las escalas de intervalos equidistantes, las de ratio también permiten el uso de técnicas de inferencia estadística.

A continuación, se describen algunos ejemplos de escalas particulares que se usan habitualmente en los cuestionarios:

Nota

Los ejemplos sólo cubren algunas de las tipologías de escalas más usadas. En Internet es fácil encontrar ejemplos de cuestionarios completos y otros tipos de escalas sin más que buscar por términos como *survey examples*, *questionnaire examples*, etc.

- La **escala de Likert**: esta escala suele usarse para obtener el grado de acuerdo o desacuerdo del encuestado con una determinada afirmación (figura 3). Puesto que todas las categorías en una escala de Likert suelen estar etiquetadas (y las etiquetas o identificadores de cada categoría no tienen por qué representar magnitudes equidistantes), hay cierta discrepancia entre los expertos sobre si esta escala debe considerarse simplemente como una escala ordinal o bien puede incluso considerarse como una escala de intervalos. Una posible solución a este problema sería mantener únicamente los identificadores o etiquetas de los extremos (p. ej.: "(1) Muy en desacuerdo" y "(5) muy de acuerdo"), dejando el resto de ítems numerados pero sin etiquetar, de modo que los números definan intervalos equidistantes. En todo caso, es éste un tema bastante discutible sobre el que no parece haber un total consenso. Obviamente, resulta muy ventajoso poder considerar una escala de Likert como de intervalos para poder así aplicar técnicas de inferencia estadística de forma lícita.

Figura 3. Ejemplo de preguntas usando una escala de Likert

```
Selecciona un número de la escala para expresar en qué medida
estás en acuerdo o en desacuerdo con cada una de las afirma-
ciones siguientes referidas a la asignatura Estadística:

                         Escala
                 1 Totalmente de acuerdo
                      2 De acuerdo
                       3 Neutral
                     4 En desacuerdo
                5 Totalmente en desacuerdo

Los exámenes finales son coherentes con la EC        _____
La asignatura ofrece contenidos prácticos            _____
Los materiales docentes son adecuados                _____
```

- La **escala de frecuencia verbal**: esta escala es muy similar a la de Likert, con la diferencia de que los ítems de la escala indican con qué frecuencia se ha llevado a cabo una determinada acción (figura 4).

Figura 4. Ejemplo de preguntas usando una escala de frecuencia verbal

```
Selecciona un número de la escala para expresar la frecuencia
con la que ocurren cada uno de los siguientes acontecimientos
referidos a las asignaturas de la titulación que cursas:

                        Escala
                      1 Siempre
                      2 A menudo
                    3 Algunas veces
                     4 Casi nunca
                       5 Nunca

Los exámenes finales son coherentes con la EC        _____
Las asignaturas ofrecen contenidos prácticos         _____
Los materiales docentes son adecuados                _____
```

- La **escala comparativa**: a diferencia de las anteriores, los ítems de esta escala indican cómo se comparan dos elementos entre sí a criterio del encuestado (figura 5). Esta escala se considera como una escala de intervalos, por lo que es lícito aplicar las técnicas de inferencia a los datos obtenidos con ella.

Figura 5. Ejemplo de uso de una escala comparativa

```
Selecciona un número de la escala para expresar tu opinión so-
bre cada uno de los siguientes temas:

                         Escala
                    1 Muy superior
                      2 Superior
                      3 Similar
                      4 Inferior
                    5 Muy inferior

Comparado con el plan de estudios anterior,
el nuevo plan de estudios te parece                      _____

Comparado con el sistema de evaluación anterior,
el nuevo sistema de evaluación te parece                 _____
```

- La **escala linealnumérica**: esta escala también es similar a la de Likert, aunque los ítems extremos suelen hacer referencia al grado de importancia que asigna el encuestado a un tema y los ítems intermedios no suelen estar etiquetados (figura 6). Por esto último, se considera una escala de intervalos.

Figura 6. Ejemplo de uso de una escala lineal-numérica

```
Selecciona un número de la escala para expresar tu opinión so-
bre el nivel de relevancia de cada uno de los siguientes temas
referidos a las asignaturas que cursas:

                            Escala
   Máxima relevancia 1   2   3   4   5   6 Mínima relevancia

El uso de recursos de Internet                   _____
El uso de materiales actualizados                _____
El uso de los foros y debates                    _____
```

- La **escala de diferencias semánticas**: esta escala consiste en definir dos extremos caracterizados por adjetivos contrapuestos y, posteriormente, definir una graduación de ítems no etiquetados entre ambos (figura 7). También se considera como una escala de intervalos.

Figura 7. Ejemplo de uso de una escala de diferencias semánticas

```
En relación a la formación que recibes en esta universidad,
selecciona un valor numérico según lo próxima que esté con
respecto a cada adjetivo:

Teórica      1   2   3   4   5   6   7    Práctica
Económica    1   2   3   4   5   6   7    Cara
Actualizada  1   2   3   4   5   6   7    Desfasada
```

2. Diseño y selección de la muestra

Como ya se ha comentado en el apartado anterior, en toda encuesta hay dos tipos de errores que conviene tener presentes: (a) el error muestral, que es la diferencia entre el estimador obtenido a partir de las observaciones (p. ej., la media muestral $\bar{x}$) y el verdadero valor del parámetro poblacional (p. ej., la media poblacional μ), y (b) el sesgo o error no muestral, que engloba todos los restantes tipos de errores que pueden ocurrir durante el desarrollo y análisis de una encuesta, es decir, errores en el diseño de las preguntas, errores causados por la "no-respuesta" (*missing data*), errores en la selección de los individuos a encuestar, errores en el registro y procesado de los datos, etc.

Ejemplo

Recordar que el término *estadístico* hace referencia a una muestra mientras que el término *parámetro* hace referencia a toda la población. Así, por ejemplo, el estadístico media muestral es un estimador del parámetro media poblacional.

Las encuestas pueden clasificarse en función del método de muestreo usado. Así, se habla de **muestreo probabilístico** cuando cada uno de los individuos que componen el marco del muestreo (elementos de la población susceptibles de ser elegidos) tiene una probabilidad conocida de ser seleccionado. Por el contrario, se habla de **muestreo no probabilístico** cuando no es posible saber cuál es la probabilidad de cada elemento de ser seleccionado. Los muestreos no probabilísticos pueden ser de gran utilidad como herramienta exploratoria, pero no permiten conocer la precisión de las estimaciones que se obtienen para los parámetros poblacionales, es decir, no dan información sobre el error muestral que se está cometiendo. Ejemplos de muestreos no probabilísticos serían los siguientes:

- A fin de conocer la opinión de los estudiantes de una universidad presencial sobre su nuevo Campus Virtual, se encuesta a los matriculados de una asignatura concreta.
- A fin de conocer la opinión de los clientes de un nuevo centro comercial, se piden voluntarios para responder a un cuestionario.
- A fin de conocer la opinión de los usuarios de una base de datos documental, un directivo selecciona una muestra de usuarios que, según su criterio, son representativos del conjunto de usuarios.

Los muestreos probabilísticos, por su parte, sí permiten calcular intervalos de confianza para los parámetros poblacionales a partir de los las observaciones de la muestra. Esto es, los muestreos probabilísticos permiten conocer la magnitud del error muestral que se está cometiendo. En este apartado se describirán cuatro de los métodos probabilísticos más populares: el muestreo aleatorio simple, el muestreo sistemático, el muestreo estratificado, y el muestreo por conglomerados.

2.1. Muestreo aleatorio simple

En un **muestreo aleatorio simple**, todos los elementos del marco muestral (elementos de la población que son candidatos a ser seleccionados) tienen la misma probabilidad de ser elegidos. Para seleccionar, mediante muestreo aleatorio simple, n elementos de entre los N que componen la lista de candidatos a ser elegidos, se suele asignar un número natural (1, 2, 3, …, N) a cada uno de los elementos de la lista y, a continuación, se generan al azar n números aleatorios distintos, que identificaran a los elementos seleccionados.

De acuerdo con la teoría de la estadística inferencial, si se selecciona una muestra aleatoria suficientemente grande (en la práctica $n \geq 30$ suele ser suficiente), el teorema central del límite permite obtener intervalos de confianza para la media poblacional μ. En particular:

Para un nivel de confianza del 95%, un intervalo de confianza para la media poblacional, μ, viene dado por:

$$\bar{x} \pm 1,96 \cdot \sqrt{\frac{N-n}{N}}\left(\frac{s}{\sqrt{n}}\right)$$

donde s representa la desviación estándar de las observaciones muestrales.

Ejemplo: un periódico de economía tiene actualmente $N = 8.000$ lectores suscritos. Una muestra aleatoria simple de $n = 484$ lectores es elegida para realizar una encuesta. Tras analizar los datos de dicha encuesta, se sabe que la media de los ingresos mensuales de los lectores seleccionados en la muestra es de $\bar{x} = 30.500$ euros y que la correspondiente desviación estándar es de $s = 7.040$ euros.

La media muestral, $\bar{x}$, es un buen estimador de la media poblacional, μ. Además, un intervalo de confianza del 95% para dicha media poblacional será: $30.500 \pm 1,96 \cdot \sqrt{\frac{8.000-484}{8.000}}\left(\frac{7.040}{\sqrt{484}}\right)$ = (29.892,07, 31.107,93). En otras palabras, para un nivel de confianza del 95%, los ingresos medios del conjunto de los 8.000 lectores suscritos al periódico oscilarán entre 29.892 y 31.108 euros.

De forma similar, es posible calcular intervalos de confianza para otros parámetros de la población, como el total acumulado de una población, por ejemplo, la demanda total de la población, la riqueza total de una población, etc.

El estadístico $N \cdot \bar{x}$ es un buen estimador del total acumulado de una población, $N \cdot \mu$. Además, si (a, b) es un intervalo de confianza del 95% para la media poblacional, μ, un intervalo de confianza del 95% para $N \cdot \mu$, viene dado por $(N \cdot a, N \cdot b)$.

Ejemplo: se desea estimar el número total de visitas anuales que reciben los portales web de las universidades pertenecientes a una clasificación que incluye las quinientas mejores del mundo. Para ello, se ha seleccionado una muestra aleatoria de cincuenta universidades pertenecientes a esa clasificación y se han obtenido los siguientes estadísticos muestrales: el número medio de visitas anuales es de veintidós mil, siendo la desviación estándar de cuatro mil.

En primer lugar, cabe destacar que $N \cdot \overline{x}$ = 11.000.000 será un buen estimador para el número total de visitas anuales que reciben los portales de las quinientas mejores universidades. Un intervalo de confianza del 95% para el número total de visitas anuales será: $500 \cdot 22.000 \pm 1{,}96 \cdot 500 \cdot \sqrt{\frac{500-50}{500}}\left(\frac{4.000}{\sqrt{50}}\right)$ = (10.474.077, 11.525.923). En otras palabras, para un nivel de confianza del 95%, el número total de visitas anuales que recibirán los quinientos portales web estará entre 10,47 millones y 11,53 millones.

Indicación

Al realizar los cálculos, se recomienda usar al menos cuatro decimales para no perder demasiada precisión en el redondeo, especialmente cuando *N* es un número muy grande.

Finalmente, también es posible obtener intervalos de confianza para la proporción de elementos de una población que satisfacen unas determinadas condiciones, por ejemplo, proporción de individuos que usan un servicio, proporción de individuos con estudios superiores, etc.

Para un nivel de confianza del 95%, un intervalo de confianza para la proporción *p* de elementos de una población que cumple una determinada condición viene dado por:

$$p' \pm 1{,}96 \cdot \sqrt{\left(\frac{N-n}{N}\right) \cdot \left(\frac{p'(1-p')}{n-1}\right)}$$

donde *p'* es la proporción de elementos de la muestra que la cumplen.

Ejemplo: siguiendo con el ejemplo anterior de los portales web de las universidades pertenecientes a la clasificación de las 500 mejores, se desea estimar el porcentaje de portales que disponen de un programa institucional –al estilo del MIT OpenCourseWare– para ofrecer contenidos formativos en abierto. De las cincuenta universidades que constituyen la muestra, un total de treinta y cinco disponen de dicho programa.

La proporción muestral, *p*' = 35/50 = 0,70 = 70%, es un buen estimador del porcentaje de universidades en las quinientas mejores que tendrán un programa así. Además, es posible obtener un intervalo de confianza del 95% para dicha proporción poblacional: $0{,}7 \pm 1{,}96 \cdot \sqrt{\left(\frac{500-50}{500}\right) \cdot \left(\frac{0{,}7(1-0{,}7)}{50-1}\right)}$ = (0,5783, 0,8217). En otras palabras, con un nivel de confianza del 95% se puede afirmar que entre el 58% y el 82% de universidades entre las quinientas mejores disponen de un programa de contenidos formativos en abierto. Observar que, en

este caso, el intervalo de confianza es poco preciso (hay unos veinticuatro puntos porcentuales de diferencia entre los extremos del intervalo), lo cual se debe a que el tamaño de la muestra es relativamente pequeño.

2.2. Muestreo sistemático

El **muestreo sistemático** consiste en usar una regla para seleccionar de forma sistemática los elementos de una muestra. Este muestreo se suele usar en poblaciones grandes y homogéneas como alternativa al muestreo aleatorio simple, especialmente en aquellas situaciones en las que el proceso de asignar un número entero a cada elemento de una larga lista puede resultar complicado o costoso en tiempo (p. ej.: asignar un número entero a cada uno de los números de una guía telefónica, asignar un número entero a cada uno de los clientes que accede a un centro comercial en un día determinado, etc.). Así, por ejemplo, si se desea seleccionar una muestra de treinta teléfonos de la guía telefónica de una ciudad, una forma sistemática de hacerlo sería escoger al azar el primero y, posteriormente, escoger un teléfono cualquiera de cada una de las veintinueve páginas siguientes. Otro ejemplo: si se desea entrevistar a cuarenta clientes de un gran centro comercial, una forma sistemática de seleccionar la muestra sería empezar por uno al azar y, a continuación, escoger cada cinco minutos al nuevo cliente que acceda al centro en ese preciso instante. A menudo, este tipo de muestreo se puede considerar como equivalente a un muestreo aleatorio simple, especialmente cuando el listado o marco muestral sigue un orden aleatorio, es decir, realizar una selección sistemática de elementos en una lista que sigue un orden aleatorio es técnicamente equivalente a realizar directamente una selección aleatoria de elementos que no sigan un orden aleatorio.

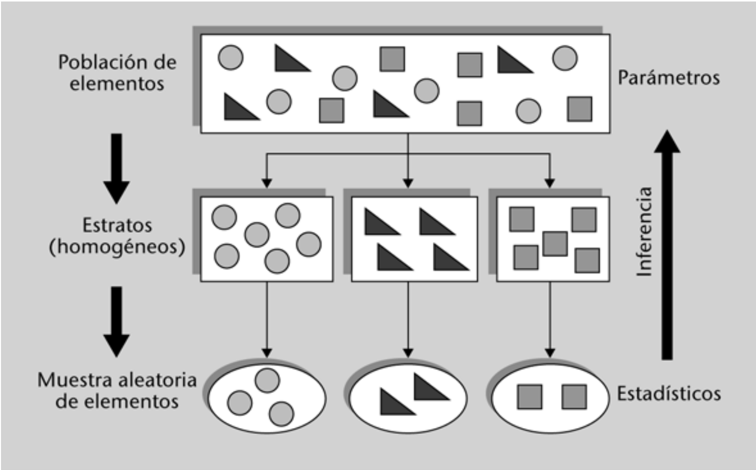
2.3. Muestreo aleatorio estratificado (grupos homogéneos)

El **muestreo aleatorio estratificado** se suele usar en los casos en que resulta fácil agrupar los elementos de la población considerada en subgrupos de composición homogénea llamados **estratos**. Por ejemplo: trabajadores de una organización agrupados por departamento, estudiantes de una universidad agrupados por titulación, habitantes de un país agrupados por nivel de renta o edad, revistas científicas agrupadas por ámbito temático, etc. Cuando la variabilidad dentro de cada estrato es menor que la variabilidad entre estratos, este tipo de muestreo tiende a proporcionar más precisión que un muestreo aleatorio simple a la hora de estimar los parámetros poblacionales.

Así, el muestreo aleatorio por estratos consiste en: (a) clasificar los N elementos de una población en *H* grupos o estratos (de manera que los elementos de cada estrato sean similares entre ellos), y (b) seleccionar a continuación una muestra aleatoria simple para cada uno de los estratos (figura 8). Los estadísticos obtenidos para cada estrato se combinan posteriormente para obtener es-

timaciones de algunos parámetros como la media, el total acumulado o la proporción de la población.

Figura 8. Muestreo aleatorio estratificado



En un muestreo por estratos, es posible obtener un buen estimador de la media poblacional haciendo un promedio ponderado de las medias muestrales obtenidas en cada estrato. En concreto, $\overline{x}_E = \frac{1}{N}\sum_{i=1}^{H} N_i \cdot \overline{x}_i$ es un buen estimador de μ, donde N_i representa el número total de elementos del estrato i-ésimo y $\overline{x}_i$ representa la media de la muestra asociada a dicho estrato.

Para un nivel de confianza del 95%, un intervalo de confianza para la media poblacional, μ, viene dado por:

$$\overline{x}_E \pm 1{,}96 \cdot \sqrt{\frac{1}{N^2}\sum_{i=1}^{H} N_i\left(N_i - n_i\right)\frac{s_i^2}{n_i}}$$

donde n_i y s_i representan, respectivamente, el tamaño y la desviación estándar de la muestra asociada al estrato i-ésimo.

Por otro lado, el estadístico $N \cdot \overline{x}_E$ es un buen estimador del total acumulado de una población, $N \cdot \mu$. Además, si (a, b) es un intervalo de confianza del 95% para la media poblacional, μ, un intervalo de confianza del 95% para $N \cdot \mu$, viene dado por $(N \cdot a, N \cdot b)$.

Finalmente, un intervalo de confianza para la proporción p de elementos de una población que cumple una determinada condición viene dado por:

$$p'_E \pm 1{,}96 \cdot \sqrt{\frac{1}{N^2}\sum_{i=1}^{H} N_i\left(N_i - n_i\right) \cdot \left(\frac{p'_i\left(1 - p'_i\right)}{n_i - 1}\right)}$$

donde $p'_E = \frac{1}{N}\sum_{i=1}^{H} N_i \cdot p'_i$ es un promedio ponderado de las proporciones p'_i de elementos de la muestra que la cumplen para el estrato i-ésimo.

Ejemplo: hace dos años se graduaron en una universidad un total de 1.500 estudiantes. Para conocer el salario medio de dichos estudiantes, tanto a nivel global como por titulación, se agruparon los estudiantes por titulaciones (estratos) y se encuestó a un total de ciento ochenta exestudiantes. La tabla 1 incluye, por orden de columnas, el número de graduados en cada estrato, el tamaño de cada muestra, la media muestral, la desviación estándar muestral y la proporción de estudiantes con un sueldo superior a los 36.000 euros anuales.

Tabla 1. Estadísticos obtenidos para cada estrato

Titulación (estrato)	N_i	n_i	$\overline{x}_i$	s_i	p'_i
Administración y Dirección de Empresas	500	45	30.000	2.000	4/45
Información y Documentación	350	40	28.500	1.700	2/40
Ingeniería Informática	200	30	31.500	2.300	7/30
Psicología	300	35	27.000	1.600	1/35
Ingeniería de Telecomunicaciones	150	30	31.000	2.250	6/30
Total	1.500	180			

Un buen estimador del salario medio para el conjunto de mil quinientos graduados viene dado por el promedio ponderado de las distintas medias muestrales:

$$\overline{x}_E = \frac{1}{1.500}\left(500 \cdot 30.000 + 350 \cdot 28.500 + 200 \cdot 31.500 + 300 \cdot 27.000 + 150 \cdot 31.000\right)$$

= 29.350 euros.

Además, se puede obtener el correspondiente intervalo de confianza, para un nivel de confianza del 95%, para la media poblacional:

$$29.350 \pm 1,96 \cdot \sqrt{\frac{1}{1.500^2}\left(500 \cdot (500-45)\frac{2.000^2}{45} + \ldots + 150 \cdot (150-30)\frac{2.250^2}{30}\right)} =$$

(29.079,33, 29.620,67), es decir, se puede afirmar, con un nivel de confianza del 95%, que el salario medio del total de mil quinientos graduados de esta universidad está entre 29.079 y 29.621 euros por año. Para hacer este tipo de cálculos es conveniente usar una hoja de cálculo (figura 9).

Figura 9. Uso de Excel para realizar cálculos en muestreo estratificado

	A	B	C	D	E	F	G	H
1	**Titulación (estrato)**	**N(i)**	**n(i)**	**x-bar(i)**	**s(i)**	**p(i)**	**N(i) * x-bar(i)**	**N(i) * (N(i) - n(i)) * (s(i)^2 / n(i))**
2	Dirección de Empresas	500	45	30.000	2.000	16.528	15.000.000	20.222.222.222
3	Información y Documentación	350	40	28.500	1.700	14.642	9.975.000	7.839.125.000
4	Ing. Informática	200	30	31.500	2.300	40.024	6.300.000	5.995.333.333
5	Psicología	300	35	27.000	1.600	12.785	8.100.000	5.814.857.143
6	Ing. Telecomunicación	150	30	31.000	2.250	39.994	4.650.000	3.037.500.000
7	*Totales*	*1.500*	*180*				*44.025.000*	*42.909.037.698*
8								
9			z =	1,96				
10			x(E) =	29.350			$\sum_{i=1}^{5} N_i \cdot \bar{x}_i$	$\sum_{i=1}^{5} N_i (N_i - n_i) \frac{s_i^2}{n_i}$
11			s(E) =	138,10				
12			x(E) - z*s(E) =	29.079,3306				
13			x(E) + z*s(E) =	29.620,6694				
14								

En segundo lugar, se pueden estimar los ingresos anuales totales del conjunto de los mil quinientos graduados, $N \cdot \mu$, para saber cuál será su potencial impacto sobre la economía local. En este caso, puesto que el estimador de μ era $\bar{x}_E$ = 29.350 euros, el estimador puntual de $N \cdot \mu$ será 1.500 $\bar{x}_E$ = 44.025.000 y un intervalo de confianza al 95% vendrá dado por: (1.500 · 29.079,3306, 1.500 · 29.620,6694) = (43.618.995,86, 44.431.004,14). En otras palabras, se puede afirmar con un nivel de confianza del 95% que serán necesarios entre 43,6 y 44,4 millones de euros para cubrir los salarios anuales de los mil quinientos graduados.

En tercer lugar, un buen estimador del porcentaje de estudiantes de la población cuyos ingresos superan los 36.000 euros vendrá dado por el promedio ponderado de los porcentajes en cada estrato: $p'_E = \frac{1}{1.500}\left(500\frac{4}{45} + \ldots + 150\frac{6}{30}\right) = 0{,}0981$, es decir, aproximadamente, sólo un 9,8% de los salarios de los mil quinientos graduados será superior a los 36.000 euros anuales. Finalmente, se puede obtener un intervalo de confianza del 95% para el porcentaje poblacional anterior:

$$0{,}0981 \pm 1{,}96 \cdot \sqrt{\frac{1}{1.500^2}\left(500(500-45)\frac{(4/45)(41/45)}{45-1} + \ldots + 150(150-30)\frac{(6/30)(24/30)}{30-1}\right)}$$

= (0,0584, 0,1379), es decir, se puede afirmar con un 95% de confianza que el porcentaje de graduados en la promoción de hace dos años cuyos ingresos superan los 36.000 euros anuales oscila entre un 5,8% y un 13,8%.

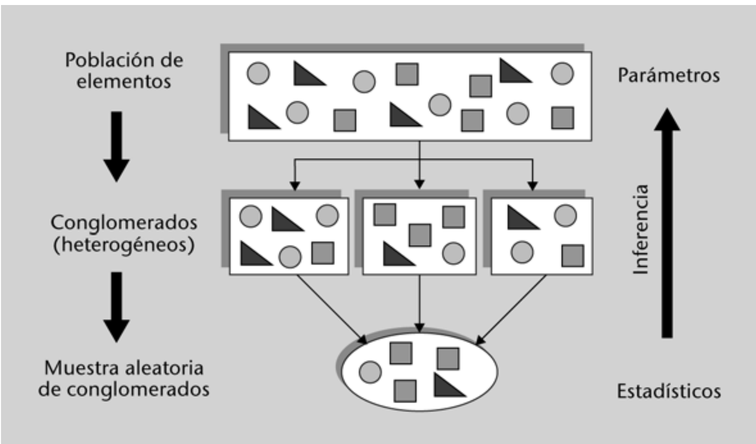
2.4. Muestreo por conglomerados (*clusters* o grupos heterogéneos)

El **muestreo por conglomerados** se suele usar en los casos en que resulta fácil agrupar los elementos de la población considerada en subgrupos de

composición heterogénea llamados conglomerados, cada uno de los cuales viene a ser una representación a pequeña escala de la población total (es decir, se presupone una gran variabilidad entre los elementos de un mismo conglomerado). Por ejemplo: los habitantes de una gran ciudad pueden agruparse por barrios, los usuarios de un servicio web pueden agruparse por países de procedencia, las revistas científicas pueden agruparse por editorial, etc. De hecho, una de las principales aplicaciones del muestreo por conglomerados está relacionada con el muestreo por áreas o regiones geográficas, donde los conglomerados suelen ser países, regiones, ciudades o barrios. El muestreo por conglomerados permite reducir los costes de desplazamientos entre zonas geográficamente dispersas y, a la vez, evita tener que generar listados exhaustivos de toda la población, puesto que sólo son necesarios listados exhaustivos de cada conglomerado seleccionado.

Así, el muestreo por conglomerados consiste en: (a) clasificar los N elementos de una población en H grupos o conglomerados (de manera que los elementos de cada conglomerado presenten mucha variabilidad entre ellos), (b) seleccionar a continuación una muestra aleatoria simple de h conglomerados, y (c) para cada conglomerado de la muestra seleccionada, o bien encuestar a cada uno de los elementos que lo componen –muestreo por conglomerados en una etapa– o bien seleccionar una nueva muestra aleatoria de elementos para encuestar –muestreo en dos etapas– (figura 10). Si bien tanto en un caso como en otro es posible obtener estimadores puntuales y por intervalos para varios parámetros poblacionales, se tratará sólo el muestreo por conglomerados en una etapa (es decir, se supondrá que, una vez seleccionada la muestra de conglomerados, se encuesta a todos los elementos de cada conglomerado seleccionado).

Figura 10. Muestreo por conglomerados



En un muestreo por conglomerados es posible obtener un buen estimador de la media poblacional μ mediante la expresión $\overline{x}_C = \dfrac{\sum_{i=1}^{h} y_i}{\sum_{i=1}^{h} N_i}$, donde N_i represen-

ta el número total de elementos del conglomerado *i*-ésimo e y_i representa el valor total de las observaciones de dicho conglomerado.

Para un nivel de confianza del 95%, un intervalo de confianza para la media poblacional, μ, viene dado por:

$$\overline{x}_C \pm 1,96 \cdot \sqrt{\frac{H-h}{H \cdot h\left(\frac{N}{H}\right)^2}\left(\frac{\sum_{i=1}^{h}\left(y_i - \overline{x}_C \cdot N_i\right)^2}{h-1}\right)}$$

Por otro lado, el estadístico $N \cdot \overline{x}_C$ es un buen estimador del total acumulado de una población, $N \cdot \mu$. Además, si (a, b) es un intervalo de confianza del 95% para la media poblacional μ, un intervalo de confianza del 95% para $N \cdot \mu$ viene dado por $(N \cdot a, N \cdot b)$

Finalmente, un intervalo de confianza para la proporción p de elementos de una población que cumple una determinada condición viene dado por:

$$p'_C \pm 1,96 \cdot \sqrt{\frac{H-h}{H \cdot h\left(\frac{N}{H}\right)^2}\left(\frac{\sum_{i=1}^{h}\left(m_i - p'_C \cdot N_i\right)^2}{h-1}\right)}$$

donde m_i es el número de elementos del conglomerado *i*-ésimo que cumple una determinada característica y $p'_C = \frac{\sum_{i=1}^{h} m_i}{\sum_{i=1}^{h} N_i}$ es buen estimador del promedio de elementos de la población que cumplen dicha característica.

Ejemplo: el sistema sanitario de atención primaria de un país está compuesto por un total de doce mil médicos distribuidos en mil centros de atención primaria (conglomerados). Con el fin de obtener cierta información sobre la población de médicos considerada, y ante la dificultad de realizar encuestas a médicos de todos los centros, se lleva a cabo un muestreo por conglomerados en el que se seleccionan de forma aleatoria un total de diez centros de atención primaria. A continuación, se pasa una encuesta a los médicos de cada uno de los centros escogidos. La tabla 2 incluye, por orden de columnas, el identificador del centro, el número de médicos que en él trabajan, el número total de visitas asociadas con una cierta enfermedad que recibe el centro en una semana normal y el número de médicos que son mujeres.

Tabla 2. Estadísticos obtenidos para cada conglomerado de la muestra

Centro (conglomerado)	Número de médicos N_i	Total de visitas y_i	Número de mujeres m_i
CAP-01	8	320	2
CAP-02	25	1.125	8
CAP-03	4	115	0
CAP-04	17	714	6
CAP-05	7	247	1
CAP-06	3	94	2
CAP-07	15	634	2
CAP-08	4	147	0
CAP-09	12	481	5
CAP-10	33	1.567	9
Totales	**128**	**5.444**	**35**

En primer lugar, un buen estimador para el número medio de visitas semanales que recibe cada médico viene dado por: $\overline{x}_C = \frac{5.444}{128}$ = 42,5313, es decir, en promedio cada médico del sistema sanitario recibirá unas cuarenta y tres visitas semanales.

Es posible obtener un intervalo de confianza del 95% para dicha media poblacional:

$$42,5313 \pm 1,96 \cdot \sqrt{\frac{1.000-10}{1.000 \cdot 10\left(\frac{12.000}{1.000}\right)^2}\left(\frac{(320-42,5313\cdot 8)^2+...+(1.567-42,5313\cdot 33)^2}{10-1}\right)}$$

= 42,5313 ± 1,96 · 1,7299 = (39,14, 45,92). En otras palabras, se puede afirmar con un nivel de confianza del 95% que el promedio de visitas semanales por médico en el sistema sanitario del país está entre 39 y 46 (figura 11).

Figura 11. Uso de Excel para realizar cálculos en muestreo por conglomerados

	A	B	C	D	E
1	Centro	Número de médicos	Total de visitas	Número de mujeres	
2	(conglomerado)	N_i	Y_i	m_i	[y(i) - x(C)*N(i)]^2
3	CAP-01	8	320	2	410,06
4	CAP-02	25	1125	8	3809,20
5	CAP-03	4	115	0	3038,77
6	CAP-04	17	714	6	81,56
7	CAP-05	7	247	1	2572,39
8	CAP-06	3	94	2	1128,54
9	CAP-07	15	634	2	15,75
10	CAP-08	4	147	0	534,77
11	CAP-09	12	481	5	862,89
12	CAP-10	33	1567	9	26722,03
13	*Totales*	*128*	*5444*	*35*	*39175,97*
14					
15		z =	1,96		
16		x(C) =	42,53		
17		s(C) =	1,73		
18		x(C) - z*s(C) =	39,14		
19		x(C) + z*s(C) =	45,92		

$$\sum_{i=1}^{10}\left(y_i - \overline{x}_C \cdot N_i\right)^2$$

En segundo lugar, se pueden estimar las visitas semanales totales del conjunto de los doce mil médicos, $N \cdot \mu$, para saber cuál será su potencial impacto sobre el sistema sanitario. En este caso, puesto que el estimador de μ era $\bar{x}_C$ = 42,5313, el estimador puntual de $N \cdot \mu$ será $12.000\,\bar{x}_C$ = 510.375 y un intervalo de confianza del 95% vendrá dado por: (12.000 · 39,1.406, 12.000 · 45,9219) = (469.687,38, 551.062,62). En otras palabras, se puede afirmar con un nivel de confianza del 95% que el sistema de atención primaria del país recibirá entre 469.687 y 551.063 visitas en una semana normal.

En tercer lugar, un buen estimador del porcentaje de médicos que son mujeres vendrá dado por: $p'_C = \dfrac{35}{128}$ = 0,2734, es decir, aproximadamente el 27,3% de los médicos del sistema de atención primaria son mujeres. Finalmente, se puede obtener un intervalo de confianza del 95% para el porcentaje poblacional anterior:

$$0,2734 \pm 1,96 \cdot \sqrt{\frac{1.000-10}{1.000 \cdot 10\left(\dfrac{12.000}{1.000}\right)^2}\left(\frac{(2-0,2734\cdot 8)^2+\ldots+(9-0,2734\cdot 33)^2}{10-1}\right)}$$

= (0,2066, 0,3402), es decir, se puede afirmar con un 95% de confianza que el porcentaje de mujeres en la población de médicos de asistencia primaria oscila entre un 20,7% y un 34,0%.

3. Análisis de cuestionarios: estudio parcial de un caso

En este apartado se presenta un caso de estudio en el que se muestran ejemplos del uso de técnicas estadísticas para analizar diferentes tipos de preguntas pertenecientes a una encuesta. El objetivo de la encuesta era obtener información concreta sobre la visión (y la actitud) de las grandes empresas de una determinada comunidad autónoma respecto al fenómeno de la externalización de los servicios, sistemas y tecnologías de la información. Para ello, se diseñó una encuesta formada por varias preguntas, algunas de ellas basadas en escalas nominales y otras en escalas de intervalos equidistantes. La población objetivo de la encuesta eran los directivos de servicios, sistemas y tecnologías de la información de las empresas, con sede social en dicha comunidad autónoma, cuyo volumen de facturación o de empleados superaban unas determinadas cantidades establecidas a priori por los investigadores. Del listado completo de empresas que cumplían dichos requisitos, se seleccionó una muestra aleatoria de cien empresas y se mandó el cuestionario a los correspondientes directivos, tras lo que se obtuvo una tasa de respuesta superior al 80%. La aleatoriedad de la muestra y la alta tasa de respuesta obtenida son dos factores imprescindibles a la hora de generalizar, con ciertas garantías, los resultados de la encuesta al conjunto de la población de empresas que satisfacen las características anteriormente descritas.

Aclaración

El objetivo último de esta sección no es explicar con detalle un caso completo de análisis de una encuesta (ya que ello requeriría de un módulo entero), sino proporcionar ejemplos concretos de cómo se pueden utilizar muchos de los conceptos y técnicas vistas en módulos anteriores para analizar encuestas. Así pues, esta sección muestra cómo se pueden combinar muchas de las técnicas estadísticas anteriormente vistas para extraer información a partir de los datos de una encuesta.

3.1. Ejemplo de uso de estadísticos descriptivos e intervalos de confianza

Una de las preguntas de la encuesta pedía especificar el número de trabajadores de plantilla del departamento de tecnologías de la información y la comunicación (TIC). Dicha pregunta está asociada a una variable aleatoria discreta, por lo que se pueden considerar los estadísticos descriptivos de la misma como muestra la figura 12.

Nota

Tanto los *outputs* como los gráficos de esta sección han sido generados con Minitab, usando los menús y opciones ya explicadas en módulos anteriores.

Figura 12. Estadísticos descriptivos de la variable "N.º de empleados"

```
Session

Descriptive Statistics: Empleados

Variable    N  N*   Mean  SE Mean  StDev  Minimum     Q1  Median     Q3
Empleados  77   5  58.87     8.79  77.15     6.00  14.00   37.00  65.00

Variable   Maximum
Empleados   450.00


One-Sample T: Empleados

Variable    N   Mean  StDev  SE Mean      95% CI
Empleados  77  58.87  77.15     8.79  (41.36, 76.38)
```

Recordatorio Minitab

Para obtener los estadísticos descriptivos, usar ***Stat*** > ***Basic Statistics*** > ***Display Descriptive Statistics***. Para obtener el Intervalo de Confianza usar ***Stat*** > ***Basic Statistics*** > ***1-Sample t***.

Esta pregunta fue contestada correctamente por un total de setenta y siete de los ochenta y dos directivos que respondieron la encuesta (cinco directivos dejaron

sin contestar esta pregunta). El promedio de trabajadores del departamento TIC es de cincuenta y nueve para las empresas que contestaron a la pregunta. Se observa también que el número de trabajadores en dicho departamento es muy variable, oscilando entre un mínimo de seis trabajadores y un máximo de cuatrocientos cincuenta, lo que hace pensar en diferentes niveles de externalización de los servicios y sistemas TIC. Puesto que la muestra es aleatoria, se ha podido obtener un intervalo de confianza para el promedio de trabajadores en departamentos TIC de todas las empresas de la población considerada. En este caso, usando un nivel de confianza del 95% se ha obtenido el intervalo (41,36, 76,38), es decir: con un 95% de confianza se puede afirmar que en promedio estos departamentos tienen entre 41 y 77 empleados. Asimismo, resulta posible agrupar los valores obtenidos para la variable anterior en categorías de empresas según el número de empleados en el departamento TIC, lo que permite obtener tablas y gráficos circulares para representar las frecuencias asociadas a cada tipo de empresa participante en la encuesta (figuras 13 y 14).

Figura 13. Tabla de frecuencias para cada categoría

Session

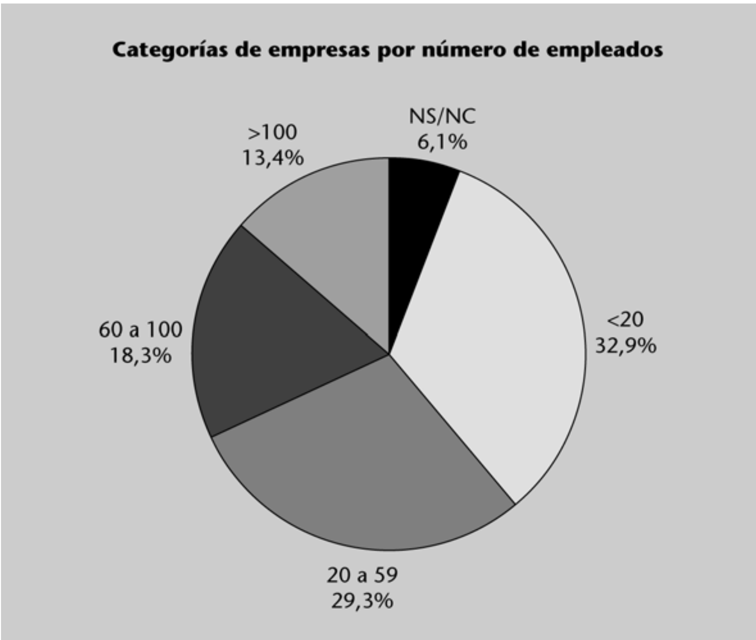
Tally for Discrete Variables: Categoría

Categoría	Count	CumCnt	Percent	CumPct
<20	27	27	32.93	32.93
>100	11	38	13.41	46.34
20 a 59	24	62	29.27	75.61
60 a 100	15	77	18.29	93.90
NS/NC	5	82	6.10	100.00
N=	82			

Recordatorio Minitab

Para obtener una tabla de frecuencias, usar ***Stat*** > ***Tables*** > ***Tally Individual Variables***.

Figura 14. Gráfico circular que representa los porcentajes de cada categoría



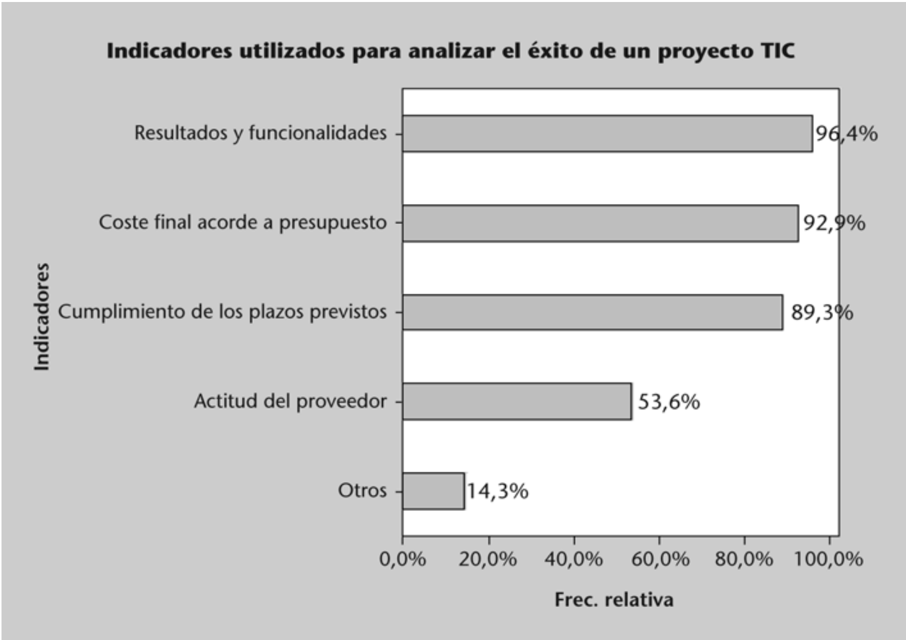
Recordatorio Minitab

Para obtener un diagrama circular, usar ***Graph*** > ***Pie Chart***.

En este caso se aprecia que aproximadamente un tercio (32,9%) de las empresas participantes tienen departamentos TIC relativamente pequeños (menos de 20 empleados), lo que induce a pensar que tendrán bastantes servicios y sistemas de información externalizados.

Otra de las preguntas de la encuesta pedía seleccionar, de entre una lista de factores, aquellos (uno o más) que se tenían en cuenta a la hora de valorar el nivel de éxito de un proyecto TIC finalizado. Puesto que se trata de una pregunta con respuesta múltiple (se pueden seleccionar varios factores a la vez), en este caso se puede emplear un diagrama de barras, como se muestra en la figura 15, para representar el porcentaje de citas de cada factor y caracterizar así aquellos factores más frecuentemente citados.

Figura 15. Gráfico de barras con frecuencia de citas de factores de éxito



Recordatorio Minitab

Para obtener un diagrama de barras, usar ***Graph*** > ***Bar Chart***. Notar que es posible personalizar los gráficos (p. ej., mostrando porcentajes, haciendo que las barras sean horizontales) mediante los botones ***Chart Options***, ***Scale***, etc. (la ayuda contextual de Minitab incluye explicaciones detalladas de todas estas opciones).

En este caso, queda claro que a la hora de valorar el éxito de un proyecto hay tres factores que se usan casi siempre ("resultados y funcionalidad", "coste final acorde a presupuesto" y "cumplimiento de los plazos previstos"). Cabe observar que el factor "otros" ha sido seleccionado en un 14,3% de las respuestas, lo que indica que tal vez exista un factor no considerado entre los anteriores que también tenga su importancia relativa.

3.2. Ejemplo de uso de contrastes de hipótesis para comparar dos grupos

En otra de las preguntas del cuestionario se le proponían al encuestado una lista de cinco ítems o motivos por los cuales una empresa podía optar por la externalización de sus servicios y sistemas TIC (p. ej.: "superar las limitaciones de las calificaciones profesionales y técnicas del equipo interno", "promover cambios organizativos, estructurales o culturales internos", "conseguir mejores niveles de calidad del servicio o sistema final", "reducir los costes totales", etc.). A continuación se le pedía valorar, usando una escala linealnumérica, la importancia de cada uno de dichos ítems o motivos de externalización, tanto desde un punto de vista teórico como desde un punto de vista práctico (es decir, el encuestado debía emitir dos evaluaciones para cada ítem: por un lado la

correspondiente a la importancia teórica o hipotética del motivo de externalización y, por otro, la correspondiente a la importancia real manifestada en la práctica cotidiana). La escala linealnumérica oscilaba entre 1 (muy poco importante) y 5 (muy importante). Uno de los objetivos de esta pregunta era determinar si para cada uno de los ítems existían diferencias significativas entre su importancia hipotética o teórica y su importancia real en la práctica del día a día (tales diferencias pondrían de manifiesto la existencia de otros factores asociados con la práctica diaria que alteraban significativamente el nivel de importancia teórico de cada motivo). En este caso se optó por realizar un contraste de hipótesis para comparar las dos medias que se obtenían para cada uno de los ítems (es decir, para cada motivo se realiza un contraste de hipótesis sobre la igualdad de la puntuación media teórica y la puntuación media práctica). La figura 16 muestra el *output* de Minitab para los dos primeros tests correspondientes a los dos primeros motivos de la lista (ítems A1 y A2). Se observa que en el caso del primer motivo de externalización considerado, no parece haber diferencias significativas, para un nivel de significación $\alpha = 0{,}05$, entre las medias respectivas de las puntuaciones teóricas (A1T) y las prácticas (A1P). Por el contrario, en el caso del segundo motivo, el *p*-valor obtenido es muy bajo (*p*-valor = 0,000), lo que evidencia la existencia de diferencias significativas entre la importancia hipotética del motivo y su importancia en la práctica.

Figura 16. Test de hipótesis para comparar medias de motivos

```
Session

Paired T-Test and CI: A1T, A1P

Paired T for A1T - A1P

              N    Mean   StDev  SE Mean
A1T          82  3.6951  0.8416   0.0929
A1P          82  3.5732  0.8017   0.0885
Difference   82  0.1220  0.8662   0.0957


95% CI for mean difference: (-0.0684, 0.3123)
T-Test of mean difference = 0 (vs not = 0): T-Value = 1.27  P-Value = 0.206


Paired T-Test and CI: A2T, A2P

Paired T for A2T - A2P

              N    Mean   StDev  SE Mean
A2T          82   3.024   1.030    0.114
A2P          82   2.707   0.923    0.102
Difference   82  0.3171  0.7180   0.0793


95% CI for mean difference: (0.1593, 0.4748)
T-Test of mean difference = 0 (vs not = 0): T-Value = 4.00  P-Value = 0.000
```

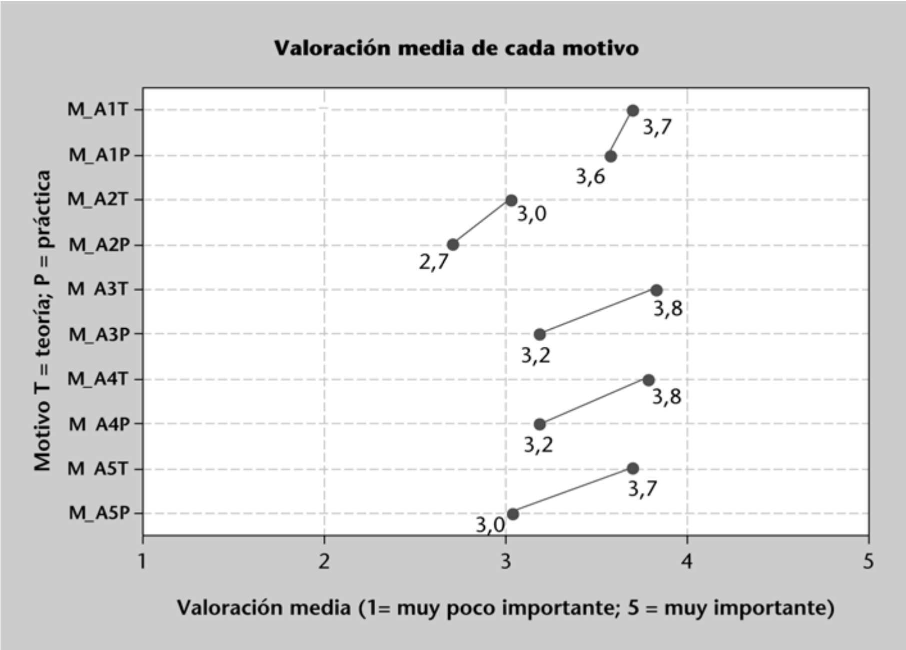
Recordatorio Minitab

Para realizar un contraste de hipótesis para dos muestras dependientes, usar ***Stat*** > ***Basic Statistics*** > ***Paired t***.

La figura 17 muestra el valor de importancia medio obtenido para cada uno de los cinco motivos de externalización considerados, tanto desde un punto de vista teórico como desde un punto de vista práctico. Se observa que, para todos los pares teoría-práctica, el valor teórico siempre es superior al valor práctico. Esto hace sospechar que si bien algunos motivos de externalización deberían ser considerados como muy importantes, en la práctica ello no siempre es posible debido a la influencia de otros factores (condi-

ciones laborales, recursos disponibles, etc.). Precisamente los contrastes de hipótesis permiten detectar aquellos casos en los que las diferencias entre teoría y práctica son significativas. Se observa también en esta figura cuál es la importancia relativa de cada motivo a la hora de decidir sobre externalizar o no los servicios y sistemas TIC.

Figura 17. Comparación visual de la importancia relativa de los ítems



Recordatorio Minitab

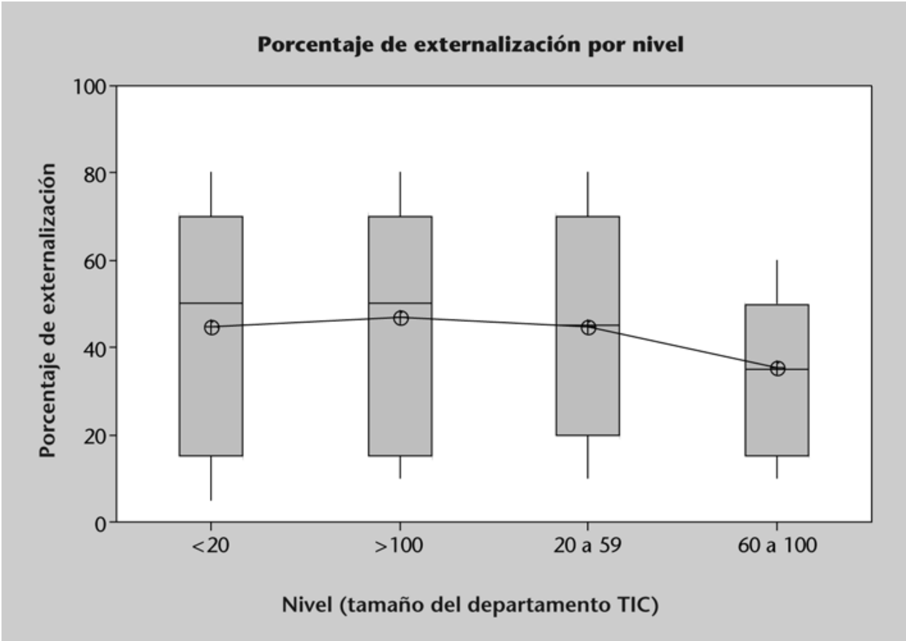
La figura muestra una nube de puntos obtenida con ***Graph*** > ***Scatterplot***. Las líneas de unión entre puntos se han generado con las opciones de dibujo de Minitab con el fin de visualizar mejor las diferencias.

3.3. Ejemplo de uso de ANOVA para comparar más de dos grupos

A fin de disponer de información sobre el porcentaje de servicios y sistemas TIC que las empresas externalizaban, en una de las preguntas se le pidió al encuestado estimar ese valor porcentual. En particular, se pretendía analizar si este porcentaje era el mismo para todas las empresas con independencia del tamaño de su departamento TIC o si, por el contrario, este porcentaje dependía de forma significativa del número de trabajadores en nómina que tuviera dicho departamento.

Puesto que se habían predefinido cuatro categorías o niveles distintos de empresas según la dimensión del departamento TIC (véase la figura 14), resulta necesario aplicar un test ANOVA para dar respuesta a la duda formulada. La figura 18 muestra una comparativa de los distintos diagramas de cajas y bigotes (*boxplots*) por categoría o nivel. Visualmente no se observan grandes diferencias entre los diferentes grupos, salvo quizá una cierta diferencia con el grupo de empresas con departamentos entre sesenta y cien empleados, cuyos porcentajes de externalización parecen algo inferiores al resto (incluso a las de mayor tamaño). En todo caso, estas posibles diferencias visuales no parecen demasiado claras.

Figura 18. *Boxplots* de porcentaje de externalización por nivel

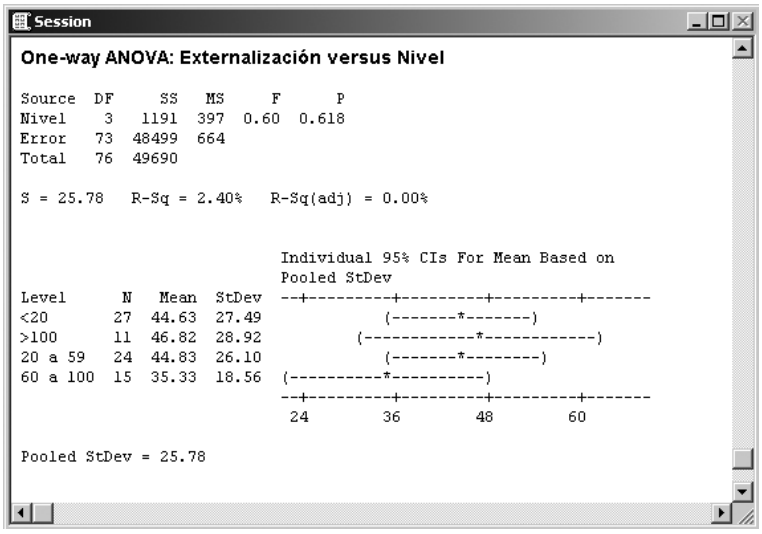


Recordatorio Minitab

Para obtener un *boxplot* múltiple, se ha de usar la opción ***Graph*** > ***Boxplot***. Las líneas de unión entre los distintos *boxplots* se generan mediante las opciones del botón ***Data View***.

La figura 19 muestra el *output* ANOVA, que ayuda a despejar las dudas: un *p*-valor de 0,618 indica que no se han hallado indicios suficientes como para rechazar la hipótesis nula de que el porcentaje medio de externalización es el mismo para todos los grupos, es decir, no parece que el tamaño del departamento TIC tenga una influencia decisiva en el porcentaje de servicios y sistemas TIC que acaban externalizándose.

Figura 19. Contraste ANOVA para comparar las medias de porcentajes


```
Session

One-way ANOVA: Externalización versus Nivel

Source  DF     SS   MS     F      P
Nivel    3   1191  397  0.60  0.618
Error   73  48499  664
Total   76  49690

S = 25.78   R-Sq = 2.40%   R-Sq(adj) = 0.00%


                             Individual 95% CIs For Mean Based on
                             Pooled StDev
Level      N   Mean  StDev   --+---------+---------+---------+-------
<20       27  44.63  27.49             (-------*-------)
>100      11  46.82  28.92          (------------*------------)
20 a 59   24  44.83  26.10             (-------*--------)
60 a 100  15  35.33  18.56   (----------*----------)
                             --+---------+---------+---------+-------
                              24        36        48        60

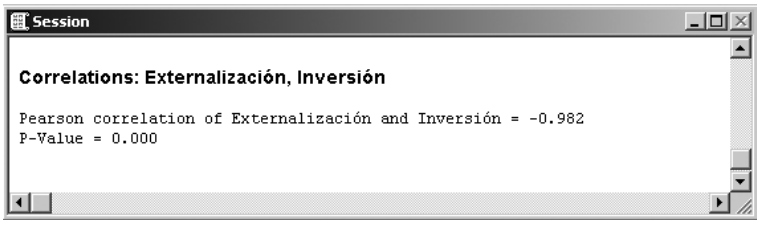
Pooled StDev = 25.78
```

3.4. Ejemplo de uso de correlación y regresión lineal

En una de las últimas preguntas del cuestionario se pedía a los encuestados estimar las cantidades (en euros) que tenían previsto invertir durante el próximo año en adquisición de programas y nuevos sistemas informáticos. Parece lógico pen-

sar que estas cantidades pueden estar inversamente relacionadas con los porcentajes de externalización de cada empresa, esto es, cabría esperar que a mayor porcentaje de externalización de servicios y sistemas TIC, menor inversión prevista en adquisición de programas y nuevos sistemas informáticos. Para tratar de corroborar esta impresión y detectar una posible correlación lineal entre ambas variables se calculó el coeficiente de correlación lineal entre ambas. La figura 20 muestra que, en efecto, existe una fuerte correlación lineal negativa entre ambas variables, ya que el coeficiente de correlación es de –0,982.

Figura 20. Coeficiente de correlación lineal entre externalización e inversión

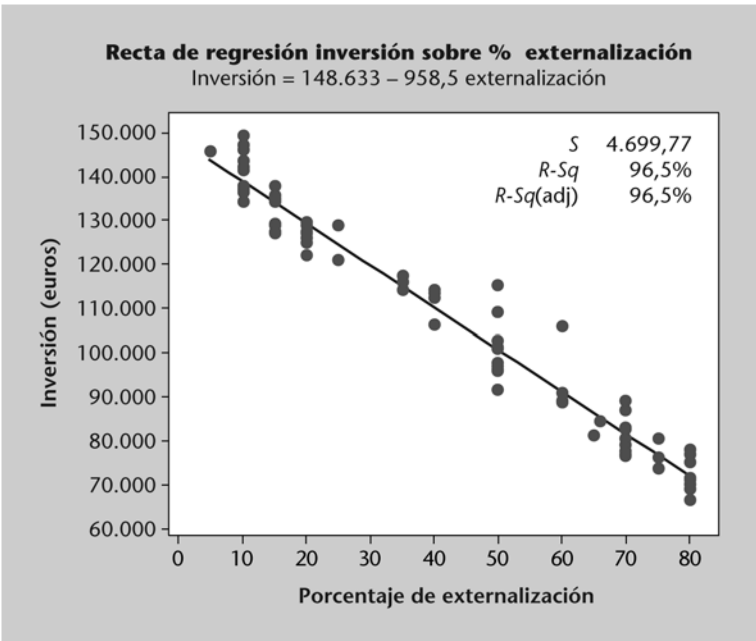


Recordatorio Minitab

Para calcular el coeficiente de correlación, usar la opción ***Stat > Basic Statistics > Correlation***.

Tiene sentido, pues, representar la recta de regresión de la inversión sobre el nivel de externalización. Esta recta se muestra en la figura 21. Puesto que el coeficiente de determinación asociado es muy alto (*R–sq* = 96,5%), se puede incluso usar al ecuación de dicha recta para hacer estimaciones sobre la inversión futura de las empresas en nuevos equipos informáticos a partir de su nivel de externalización de servicios y sistemas TIC.

Figura 21. Recta de regresión de la inversión sobre el nivel de externalización



Recordatorio Minitab

Para representar la recta de regresión, usar la opción ***Stat > Regression > Regression*** (alternativamente, también se puede usar ***Stat > Regression > Fitted Line Plot***).

Resumen

Las técnicas de investigación social basadas en el uso de encuestas y cuestionarios están cada vez más extendidas en todos los ámbitos. Sin embargo, diseñar un buen cuestionario no es una tarea fácil y conviene tener presentes aspectos clave como la brevedad y claridad de las preguntas, el tipo de escala usada o el análisis posterior que se pretende aplicar a los datos de la muestra.

En el diseño del cuestionario y del muestreo hay que tratar de minimizar tanto el error muestral como el error no muestral o sesgo. Para ello resulta necesario conocer bien las diferentes técnicas básicas de muestreo que se usan en cada caso (muestreo aleatorio simple, muestreo sistemático, muestreo estratificado y muestreo por conglomerados).

Finalmente, una vez obtenidos los datos de la encuesta, conviene saber qué técnicas estadísticas se pueden aplicar en cada caso y qué tipo de información pueden proporcionar, tanto de forma numérica como gráfica. Precisamente, el análisis de los resultados obtenidos mediante el uso de estas técnicas comporta a menudo un proceso de reflexión importante, es decir, el programa estadístico siempre será capaz de calcular números y generar resultados, pero no siempre estos resultados tendrán sentido ni serán válidos. Es tarea del investigador comprobar si se satisfacen las hipótesis necesarias para aplicar cada técnica estadística, e interpretar y validar, si procede, los resultados generados por los ordenadores.

Ejercicios de autoevaluación

1) Seleccionar un tema y diseñar un cuestionario para obtener información sobre el mismo. El cuestionario debe contener una pregunta por cada tipo de escala (nominal, ordinal, de intervalos equidistantes y de ratio). Argumentar la validez del cuestionario y especificar qué tipo de técnicas estadísticas se pueden hacer servir para analizar cada pregunta.

2) Entre los investigadores de una universidad se ha realizado un estudio para conocer sus hábitos de trabajo. Entre otras cosas, el estudio pretendía obtener información sobre el número medio de artículos que un investigador lee anualmente, así como sobre qué porcentaje de los mismos están en inglés. Dado que la universidad tiene tres grandes ámbitos de investigación (*E-learning*, Computación y Sociedad de la Información), se diseñó un muestreo por estratos en el que se clasificó a cada investigador en el estrato correspondiente a su ámbito de investigación. La tabla siguiente resume los datos de la encuesta:

Ámbito de investigación (estrato)	N_i	n_i	x_i	s_i	p'_i
E-learning	200	20	138	30	0,50
Computación	250	30	103	25	0,78
Sociedad de la Información	100	25	210	50	0,21

Con la ayuda de un modelo de hoja de cálculo (MS Excel u Open Office Calc), se pide:

a) Obtener un intervalo de confianza del 95% para el promedio de artículos leídos anualmente por la población de investigadores de la universidad.

b) Obtener un intervalo de confianza del 95% para el total de artículos leídos anualmente por el conjunto de investigadores de la universidad.

c) Obtener un intervalo de confianza del 95% para el porcentaje de artículos leídos que están en inglés.

3) Las veinticinco bibliotecas universitarias de un país emplean un total de trescientos profesionales en su servicio de obtención de documentos (SOD). A fin de obtener información sobre el número medio de documentos "difíciles de obtener" que se solicitan anualmente, se selecciona una muestra aleatoria de cuatro bibliotecas universitarias y se encuesta a cada uno de los profesionales del SOD respectivo. También se quiere obtener información sobre el número de expertos en Tecnologías de la Información y Comunicación de cada servicio SOD analizado. La tabla inferior muestra la información obtenida:

Biblioteca (conglomerado)	Número de profesionales N_i	Total de documentos "difíciles" y_i	Número de expertos en TIC m_i
SOD-01	7	95	1
SOD-02	18	325	6
SOD-03	15	190	6
SOD-04	10	140	2

Con la ayuda de un modelo de hoja de cálculo (MS Excel o Open Office Calc), se pide:

a) Obtener un intervalo de confianza del 95% para el promedio de documentos "difíciles de obtener" procesados anualmente por un SOD.

b) Obtener un intervalo de confianza del 95% para el total de documentos "difíciles de obtener" que son procesados anualmente por el global de los SOD.

c) Obtener un intervalo de confianza del 95% para el porcentaje de especialistas en TIC que trabajan en los SOD del sistema universitario.

4) En un estudio se entrevistó a ocho individuos elegidos al azar para evaluar el potencial de venta de un producto antes y después de lanzar una fuerte campaña publicitaria por televisión. El

interés por comprar el producto fue determinado por cada individuo, antes y después de la campaña, usando una escala entre 0 y 10, donde los valores más grandes representaban un interés mayor en adquirir el producto. La tabla siguiente muestra los resultados obtenidos:

Individuo	Después	Antes
1	6	5
2	6	4
3	7	7
4	4	3
5	3	5
6	9	8
7	7	5
8	6	6

Contrastad la hipótesis nula de que, en promedio, el interés por adquirir el producto no ha variado tras la campaña. Usad un nivel de confianza del 95%.

5) En un estudio se visitaron cinco ciudades de una provincia para preguntar a los residentes sobre sus hábitos a la hora de hacer la compra. Una de las preguntas versaba sobre el número de días por mes que realizaban la compra fuera de su provincia. Un total de treinta personas participaron en la encuesta y proporcionaron las observaciones que se incluyen en la tabla siguiente:

Ciudad 1	Ciudad 2	Ciudad 3	Ciudad 4	Ciudad 5
1	3	1	2	5
3	3	6	5	3
2	4	2	7	2
1	3	5	4	9
1	9	6	8	8
0	7	3	1	6

Se pide:

a) Determinar si existen o no diferencias significativas entre los hábitos de compra de los residentes en función de su ciudad (usar un nivel de confianza del 99%).

b) Obtener el coeficiente de correlación entre la ciudad de residencia y el número de veces por mes que se compra fuera de la provincia.

Solucionario

1) Pregunta abierta, consultad el primer apartado de este material para comprobar la validez del cuestionario propuesto.

2) La figura siguiente muestra los resultados obtenidos con el modelo Excel. Así, con un nivel de confianza del 95% se puede afirmar que:
a) El número medio de artículos leídos por año e investigador oscila entre 129 y 142.
b) El total de artículos leídos por año oscila entre 70.675 y 78.025.
c) El porcentaje de los artículos leídos que está en inglés oscila entre el 47% y el 68%.

	A	B	C	D	E	F	G	H	I	J
1	**investigación (estrato)**	**N(i)**	**n(i)**	**x-bar(i)**	**s(i)**	**p'(i)**	**N(i) * x-bar(i)**	**N(i) * (N(i) - n(i)) * (s(i)^2 / n(i))**	**N(i) * p'(i)**	**N(i) * (N(i) - n(i)) * [p'(i) * (1 - p'(i)) / (n(i) - 1)]**
2	E-learning	200	20	138	30	0,50	27.600	1.620.000	100	473,68
3	Computación	250	30	103	25	0,78	25.750	1.145.833	195	325,45
4	Sociedad de la Información	100	25	210	50	0,21	21.000	750.000	21	51,84
5	*Totales*	*550*	*75*				*74.350*	*3.515.833*	*316*	*850,98*
6										
7			z =	1,96						
8			x(E) =	135,18		p'(E) =	0,57			
9			s(E) =	3,41		sp(E) =	0,05			
10		x(E) - z*s(E) =		**128,50**	p'(E) - z*sp(E) =		**0,47**			
11		x(E) + z*s(E) =		**141,86**	p'(E) + z*sp(E) =		**0,68**			
12			N * a =	**70.674,89**						
13			N * b =	**78.025,11**						
14										

3) La figura siguiente muestra los resultados obtenidos con el modelo Excel. Así, con un nivel de confianza del 95% se puede afirmar que:
a) El número medio de documentos "difíciles" solicitados por año en cada SOD oscila entre 129 y 142.
b) El total de documentos "difíciles" solicitados por año en el conjunto de los SOD oscila entre 3.635 y 5.365.
c) El porcentaje de especialistas en TIC de entre los empleados en el conjunto de los SOD oscila entre 0,21 y 0,39.

F3 fx =(D3-E10*B3)^2

	A	B	C	D	E	F
1	**SOD**	**Número de profesionales**	**Total de documentos "difíciles"**	**NNúmero de especialistas TIC**		
2	**(conglomerado)**	**N(i)**	**y(i)**	**m(i)**	**[y(i) - x(C)*N(i)]^2**	**[m(i) - p'(c)*N(i)]^2**
3	SOD-01	7	95	1	100,00	1,21
4	SOD-02	18	325	6	3025,00	0,36
5	SOD-03	15	190	6	1225,00	2,25
6	SOD-04	10	140	2	100,00	1,00
7	*Totales*	*50*	*750*	*15*	*4450,00*	*4,82*
8						
9		z =	1,96			
10		x(C) =	15,00	p'(C) =	0,3	
11		s(C) =	1,47	sp(C) =	0,05	
12		x(C) - z*s(C) =	**12,12**	p'(C) - z*sp(C) =	**0,21**	
13		x(C) + z*s(C) =	**17,88**	p'(C) + z*sp(C) =	**0,39**	
14		N*a =	**3635,18**			
15		N*b =	**5364,82**			

4) En este caso, se requiere usar un contraste de hipótesis para dos poblaciones dependientes (ya que son los mismos individuos los que contestan al test en dos momentos distintos). El *output* de Minitab muestra un p-valor = 0,217 > 0,05 = α, es decir, no se puede rechazar la hipótesis nula de que ambas medias son iguales. En otras palabras, no se han encontrado evidencias suficientes como para afirmar, a un nivel de confianza del 95%, que la campaña publicitaria ha tenido efecto en la intención de compra del producto por parte de los consumidores.

Session

Paired T-Test and CI: Antes, Después

```
Paired T for Antes - Después

             N    Mean  StDev  SE Mean
Antes        8   5.375  1.598    0.565
Después      8   6.000  1.852    0.655
Difference   8  -0.625  1.302    0.460


95% CI for mean difference: (-1.714, 0.464)
T-Test of mean difference = 0 (vs not = 0): T-Value = -1.36  P-Value = 0.217
```

5) En el *output* siguiente de Minitab se muestra un estadístico $F = 2{,}85$ que tiene un *p*-valor asociado $p = 0{,}045 > 0{,}01 = \alpha$ (puesto que en este caso el nivel de confianza era del 99%). Así pues, no hay evidencias suficientes como para rechazar la hipótesis nula de que todas las medias son iguales, es decir, no parece haber diferencias significativas entre los hábitos de compra de los residentes de las distintas ciudades. Se observa que, en efecto, los intervalos de confianza se solapan unos sobre otros.

```
Session

One-way ANOVA: Ciudad 1, Ciudad 2, Ciudad 3, Ciudad 4, Ciudad 5

Source  DF      SS     MS     F      P
Factor   4   62.00  15.50  2.85  0.045
Error   25  136.00   5.44
Total   29  198.00

S = 2.332   R-Sq = 31.31%   R-Sq(adj) = 20.32%


                           Individual 99% CIs For Mean Based on
                           Pooled StDev
Level     N   Mean  StDev  -----+---------+---------+---------+----
Ciudad 1  6  1.333  1.033  (---------*----------)
Ciudad 2  6  4.833  2.563                (---------*----------)
Ciudad 3  6  3.833  2.137            (---------*----------)
Ciudad 4  6  4.500  2.739              (----------*----------)
Ciudad 5  6  5.500  2.739                  (----------*----------)
                           -----+---------+---------+---------+----
                              0.0       2.5       5.0       7.5

Pooled StDev = 2.332
```

El *output* siguiente muestra que la correlación entre la variable Ciudad y la variable Días (ambas generadas a partir de los datos iniciales) es de 0,440, valor que no parece corresponder con una correlación fuerte. En efecto, el *p*-valor de 0,015 hace pensar que, para un nivel de confianza del 99%, ambas variables no están fuertemente correlacionadas. Observad que esta conclusión es bastante coherente con la obtenida anteriormente para el test ANOVA.

```
Session

Correlations: Días, Ciudad

Pearson correlation of Días and Ciudad = 0.440
P-Value = 0.015
```